UNIFIED LAW FOR VISUAL TASKS (THE FOUR LAWS OF COMPUTER VISION)

Sen Fang[1] **Yalin Feng**[2] **Yancheng Zhang**[3] **Yanxin Zhang**[4] **Zonglin Lyu**[3] **Hongbin Zhong**[5]
Liwei Che[1] **Rui Wu**[1] **Chen Chen**[3] **Dimitris N. Metaxas**[1]

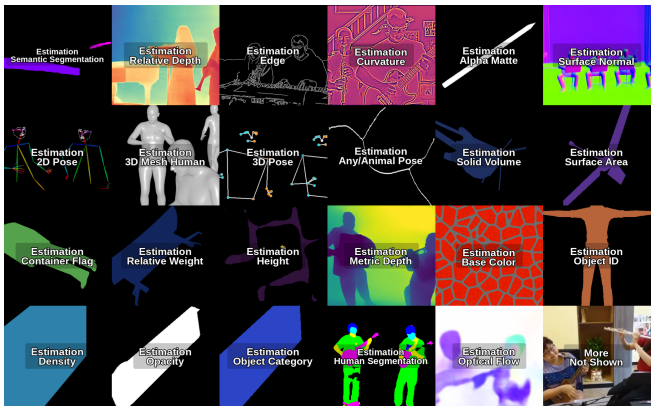
[1]Rutgers University, USA [2]NTU, Singapore [3]University of Central Florida, USA
[4]University of Wisconsin–Madison, USA [5]Georgia Institute of Technology, USA

Project Page: https://world-snapshot.github.io/WSMs

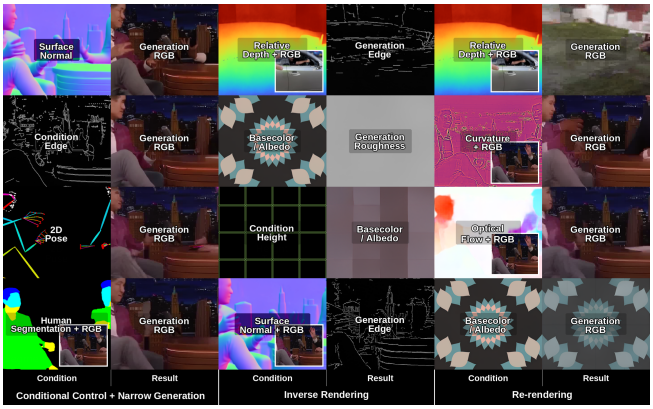
Code Dataset WebUI Document



Lex I. The previous generation was mostly generating RGB as a special case of visual generation.



Lex II. The previous estimations (understand, seg, detect) were obtaining additional channel information from RGB.



Lex III. The previous conditional control was generating RGB based on the additional channels.



Lex IV. The previous reconstruction was estimating shape information from RGB and implicitly or explicitly utilizing it.

Figure 1. **The Unified Law for Visual Task** – flattens the world into a two-dimensional space, and visual generation is almost lossless world generation. The detailed content of **The Four Laws of Computer Vision** is as shown above.

Comment: This work began in December 2024. Project page at xx; Contributions and timeline at xx; Model at xx; This is an empirical article that attempts to unify **thousands of visual tasks** through a single methodology.

Abstract

In this paper, we propose a Unified Law for Visual Tasks, which could be further divided into Four Laws of Computer Vision, with the goal of enabling a single framework or method to uniformly train on and perform nearly all visual tasks. Our starting point is an inspiration and hypothesis: if the world information is encoded into image, which we call World Snapshots, then image processing is world processing, and image generation is World Snapshot generation. Furthermore, if all visual and visualization tasks can be reformulated as generation tasks, then developing one modular model, tentatively referred to as a World Snapshot Model, would allow us to perform the vast majority of visual tasks, much as large language models have largely unified natural language processing. This paper aims to summarize and validate these laws across more than 1,000 visual tasks, and to propose a unified training framework for guiding future research and applications in visual tasks.

1. Introduction

Recently, many researchers now expect computer vision to develop a unified model as powerful as large language models [1, 2], often referring to such a model as a world model [3, 4]. This expectation is usually motivated by the hope that ideas such as "predicting the next token" or "predicting the next state" can be transferred to the visual domain [5–7]. However, this view faces two fundamental difficulties: visual data cannot be fully summarized by a simple prompt-answer format, and computer vision tasks appear in many different forms. In this paper, we therefore attempt to formulate an ideal World Model that can move beyond these limitations and improve the unity and generality of visual methods.

Insight and Hypothesis. First, we argue that what the visual domain truly needs is to encode all information into the image itself. More specifically, **if all information can be encoded into the image itself, then two-dimensional image generation becomes lossless world generation.**[1]

To verify the existence of this Unified Law, we introduce an image data format with up to 256 channels named World Snapshot Graphics (.wsg). This format stores additional information, including spatial structure, pose, physics, audio, and semantics, in extra channels. We then attempt to train a model to generate images in this format, using the generation process itself as a test of the proposed law.

Derived Implications and Research Reframing. Furthermore, a task is a visual task as long as either its input or its output is visual information. Then, visual tasks can be abstracted at a macro level into generation, estimation, conditional control, and reconstruction.[2] These four broad categories are not arbitrary: they correspond to the four fundamental directional relations formed when RGB, text, and non-RGB world attributes serve as conditions or targets. Prior work covers each category in a narrow form, such as generating only RGB or estimating a few attributes from RGB; we call this narrow vision, and its extension to every world attribute broad vision. More directly, we do not want to train four separate classes of world-processing models for generation, estimation, conditional control, and reconstruction; we treat them all as forms of generation. Figure 1 presents these four task families as Lex I–IV. Together, they refine and extend the Unified Law by treating all four as forms of generation, thereby avoiding having to train four broad classes of world snapshot/frame processing models.

The Elephant in the Room. For a concrete method, we should consider how to train a unified visual generation model whose backbone can be reused by countless conditions to perform one or multiple rounds of inference. However, this immediately raises a long-standing challenge: training efficiency. We should not substantially increase the number of parameters merely to make one model master more than 1,000 visual tasks; instead, we should aim for a regime in which the model's capabilities expand by orders of magnitude while its parameter count grows only slightly. ControlNet [11], LoRA [12], MoE [13–16], IP-Adapter [17], and IC-Control Style [18–21] are five commonly used solutions, but even LoRA, which can acquire a new capability within several hours to several days, becomes insufficient when facing thousands of tasks.

Method. We propose a new method named TunnelingNet for training WSMs. Its core idea is to assign each modality to a separate learning stream connected through a shared generative backbone. The streams are analogous to two observers, one with typical color vision and the other with red–green color blindness: each describes the same underlying object through its own representation without adopting the other's visual perception. With this method, a model can master thousands of visual tasks within a very short time. On structured non-RGB mappings, our proposed Modality Tunneling Attention (MTA) yields 19.0–38.5× update-based speedups over a parameter-matched LoRA native-attention baseline and produces the first visually identifiable Edge structure after roughly 50 updates, versus 3,850–5,000 updates for the matched NJA baseline (77–100× fewer updates; Fig. 5).

Contributions. The main contributions of this paper are

[1] This statement is the **Unified Law** proposed in this paper. Physicists at the end of the last century referred to a similar idea as the holographic principle and provided systematic arguments showing that [8–10], under special conditions, an $(n+1)$-dimensional space can be encoded into an n-dimensional space. This also supports our view that an information-rich image can approximate a world frame.

[2] These correspond to the four sub-laws. In our experiments, we find that preparing four types of interfaces can support almost all visual tasks.

summarized as follows:

- We propose **Unified Law for Visual Tasks**, which reformulates, generation, estimation, control, and reconstruction under a common generative perspective and provides a theoretical path toward unifying large-scale visual tasks.
- We introduce **World Snapshot Graphics (.wsg)**, an image format with up to 256 channels that encodes spatial, pose, physical, audio, semantic, and other world information into image-like representations for training World Snapshot Models.
- We introduce **World Snapshot Models (WSMs)**, a unified framework that organizes visual tasks into four generative interfaces, showing how different visual tasks can be treated as compatible task forms.
- We propose **TunnelingNet**, an efficient training method that independently optimizes task-specific parameters and registers them in one sparsely activated WSM around a shared visual generation backbone, enabling rapid scaling to thousands of tasks.

Contents

1 Introduction **2**

2 Related Work **6**

3 Method **6**

3.1 World Snapshot Graphics . . . 6
3.2 Four Generative Interfaces of Training . . . 7
3.3 Modality Architectures . . . 8
3.4 TunnelingNet . . . 9
3.5 Deep Tunneling Learning . . . 10
3.6 Categorical Output Head . . . 10
3.7 Sparse-Structure Modeling . . . 11

4 Ablation Evaluation **13**

4.1 Attention Topology . . . 13
4.2 Condition Routing and Tunnel Placement . . . 15
4.3 Rectified-Flow Objective . . . 17
4.4 Minute-Scale Rapid Learning . . . 17

5 Large-Scale Quantitative Evaluation **19**

5.1 Cross-Modal Matrix Analysis . . . 19
5.2 Prediction–Target Correspondence Calibration . . . 22
5.3 Three Complementary Levels of Evidence: Breadth, Genuine Task Learning, and Mechanism . . . 25
5.4 External Comparison on the AVIS Evaluation Set . . . 28
5.5 Evaluation on Post-Release Visual Data . . . 33
5.6 Evaluation with Independent Public Annotations . . . 37

6 Qualitative Evaluation **39**

6.1 Scope and Organization . . . 39
6.2 Part I: Four Inference Interfaces . . . 39
6.3 Parts II–X: Task-Family Coverage . . . 39
6.4 Interface Comparison . . . 66
6.5 RGB Estimation . . . 66
6.6 Narrow RGB Generation . . . 66
6.7 Conditional Control . . . 67
6.8 RGB-Assisted Inverse Rendering . . . 68
6.9 Categorical Output Analysis . . . 68
6.10 Sparse Digitization and Extended Tasks . . . 68

7 Discussion **69**

7.1 Experimental Evidence . . . 69
7.2 Unified Task Graph and Model Scale . . . 69
7.3 Mathematical Scalability of the Unified Law . . . 69
7.4 Unifying and Expanding Computer Vision . . . 71
7.5 Formal Scope and Falsifiability of the Unified Law . . . 72

8 Conclusion **75**

9 Appendix: Implementation Details **81**

10 Appendix: Extended Experimental Analysis **82**

10.1 Ablations and Categorical Results . . . 82
10.2 More Results . . . 83
10.3 Capability Showpiece: Speedrunning Celebrated Visual Systems . . . 84

11 Appendix: ControlWorld User Guide 86
11.1 The Four WSM Interfaces . . . 86
11.2 RGB Estimation of Geometry and Motion . . . 86
11.3 Narrow Generation from Geometry and Human Structure . . . 86
11.4 Conditional Control . . . 86
12 Appendix: Historical Task Interface Analysis 103
12.1 Novelty of evaluated WSM tasks . . . 103
12.2 Historical Tasks Covered by the Four Interfaces . . . 104
13 Appendix: Tutorial 171
13.1 Background: Two AI Fields, Two Different Outcomes . . . 171
13.2 Three Goals of Visual-Task Unification . . . 171
13.3 Why Large-Scale Visual Unification Remains Difficult . . . 171
13.4 Current Methods and a Sixth Paradigm . . . 172
13.5 A Detailed Comparison of the Six Paradigms . . . 174
13.6 What the Comparison Means . . . 176
13.7 What Constitutes the Unified Capability . . . 176
13.8 How Parameter and Training Costs Scale with More Capabilities . . . 176
13.9 Capabilities That Emerge at the System Level . . . 177
13.10 A Practical Mental Model of WSG, WSM, and TunnelingNet . . . 178
13.11 From a Route to a Visual Capability Operating System . . . 178
13.12 Explicit Workflows and an Embedded Task Graph . . . 179
13.13 World Snapshot Model Application Landscape . . . 180
14 Memoir: Experimental Derivation of the Unified Law 186
14.1 We First Tried to Put All World Information into One Codec . . . 186
14.2 Early Sparse Codecs Exposed a Representation Problem . . . 187
14.3 We Then Built Small Bridges between Latent Spaces . . . 188
14.4 We Tested Conversion inside Pretrained Video Generators . . . 190
14.5 We Moved the Conversion into the Generation Process . . . 191
14.6 We Extended the Framework to New Target Structures . . . 192
14.7 The Experiments That Led to the Unified Law . . . 194
14.8 How the Experiments Led to the Unified Law . . . 209
14.9 The Deeper Insight: a Model Must First Preserve What It Is Asked to Learn . . . 209
14.10 What the Experiments Taught Us about Modalities . . . 210
14.11 Formal Lessons from the Experiments . . . 212
14.12 From Four Initial Tunnels to 1,148 Trained Baselines . . . 225
14.13 Summary of the Unified Law . . . 225

Table 1. Comparison between World Snapshot Models (WSMs) and Large Language Models (LLMs).

Comparison Aspect	WSMs	LLMs
Domain	Vision	NLP
Methodology	Encode Information into Image	Predict Next Token
Main Laws/Principles Realized	**Unified Laws**: all vision tasks are generative forms	**Scaling Laws**: emergent intelligence, prompt-answer pairs
Can Unify All Corresponding Tasks	Yes	Yes
Can Subsume Other Domains	Yes	Yes
Can Scale Domains and Tasks Multiplicatively	Yes	Yes
Zero-Shot/Training-Free/Plug-and-Play/Scalable	Yes	Yes
Transcends Original Domain as Almost Entirely New Research	Yes	Yes
Chain of Thought/Visual[3] (CoT/CoV)	Yes	Yes

Note. According to the Papers with Code data, there are 1477 commonly recorded Vision tasks and 858 commonly recorded NLP tasks.

2. Related Work

Physical Holographic Principle [8–10]. Proposed in theoretical physics, it suggests that all the information contained within a volume of space can be encoded on its boundary surface. This principle originated from black hole thermodynamics, where the entropy S of a black hole is proportional to its surface area A rather than its volume: $S = \frac{k_B A}{4l_p^2}$ where k_B is the Boltzmann constant and l_p is the Planck length. This implies that a three-dimensional universe might be described by information living on its two-dimensional boundary. This principle suggests that the complete universe can be fully represented by a lower-dimensional hologram [8]. Put differently, a special image can be no different from a world in one dimension higher, and they are inherently the same. Our approach originated from our own inspiration, but it coincides with this theory. To a large extent, it supports our methodology.

Rectified Flow Models [22–26]. Rectified Flow presents a simple and efficient approach to learning ordinary differential equation (ODE) models for transporting distributions between two empirically observed distributions π_0 and π_1 [27–29]. The core idea is to learn ODEs that follow straight paths connecting points sampled from the source and target distributions as much as possible. Given samples $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$, the method constructs a linear interpolation:

$$X_t = (1-t)X_0 + tX_1, \quad t \in [0,1]$$

and trains a neural network to predict the velocity field $v(x,t)$ by minimizing the straightness loss:

$$\mathcal{L} = \mathbb{E}_{t,X_0,X_1}\left[\|v(X_t,t) - (X_1 - X_0)\|^2\right]$$

The learned velocity field defines an ODE that transports the distribution: $\frac{\mathrm{d}Z_t}{\mathrm{d}t} = v(Z_t,t), \quad Z_0 \sim \pi_0$ Straight paths are preferred because they minimize discretization error and enable fast inference with very few Euler steps. The method supports a "reflow" procedure that iteratively straightens trajectories, ultimately achieving high-quality generation with a single discretization step. The rectified flow is the main reason why generative models have become more powerful in the past two years. Unlike the classical diffusion [30], this model predicts the velocity field along the transport trajectory.

Multimodal Joint Generation. Our method builds on the generative formulation discussed above, since RF/FM-based DiTs have recently become a common backbone for more powerful visual generation systems: Open-source image and video generation models such as Flux [26, 31], Wan-Video [32], and LTX [33, 34] have first advanced the capability of visual generation. More recently, models such as Google Veo [35], OminiControl [36, 37], Vision Banana [38], and UniVidX [39] have attempted large-scale multimodal joint generation. These generative methods can be viewed as attempts to pursue a goal similar to ours; however, due to resource constraints, their exploration is still not sufficiently thorough or deep.[4]

3. Method

3.1. World Snapshot Graphics

We introduce World Snapshot Graphics (.wsg) as a unified image format for representing World Snapshots. A .wsg sample is an image-like representation addressed by up to 256 channel IDs, where standard visual appearance occupies part of the representation and auxiliary channels encode richer world attributes. The 256 channels are a management convention: channel IDs form an address space currently capped at 256, and string names would serve equally well. Each task reads and generates only its own

[3]The Chain of Visual refers to the process of transferring within the visual network we have constructed, such as ordinary transfers; complex transfers involving the addition or reduction of input types; and the process of reasoning based on visual evidence (For example, the color brown can be associated with tree bark, roughness, and hardness.), etc.

[4]According to timelines recorded by third-party systems, we began exploring this direction two years earlier than related work, and our exploration is substantially deeper, allowing our public papers to cover roughly 50–150 times more tasks.

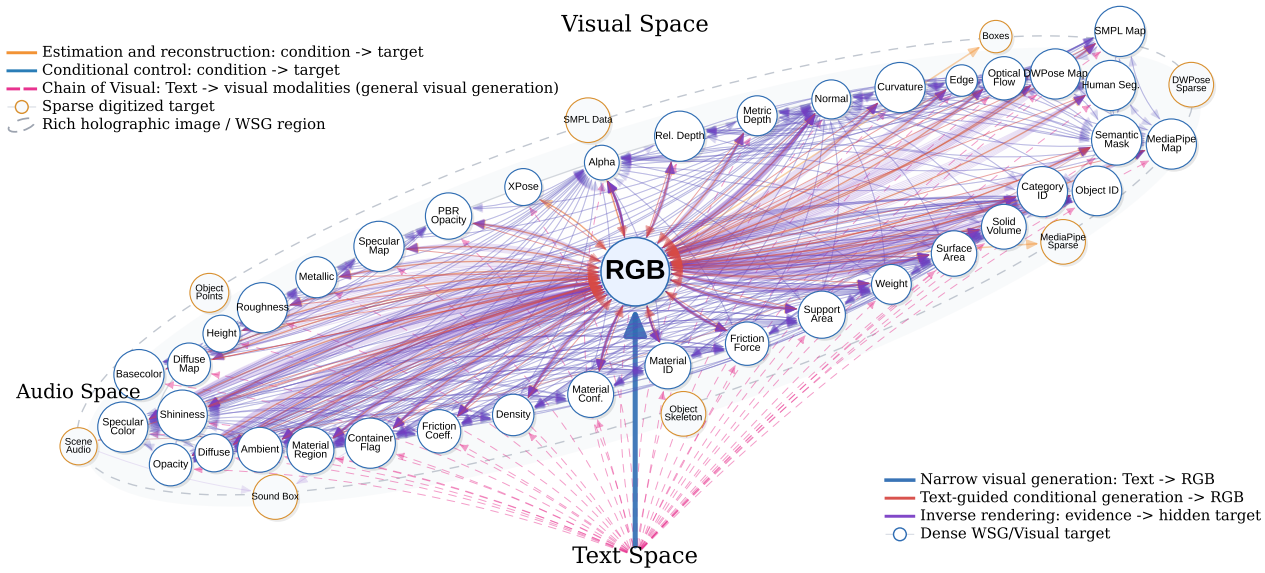


Figure 2. **WSG task graph and the four generative interfaces.** RGB is the center of the visual task space. Text enters from the bottom and generates RGB; WSG modalities surround RGB as target or condition nodes. Each directed arrow is a trainable Channel Head, also called a Tunnel Head: a compact network parsed from the WSG modality graph and trained through TunnelingNet to steer the backbone's generative trajectory. Arrow color indicates the task type: visual generation, estimation/reconstruction, conditional control, text-guided conditional generation, or inverse rendering. Dashed arrows denote Chain of Visual, where text can reach non-RGB World Snapshot modalities through generated World Snapshots. Blue-outline nodes are dense WSG/visual targets, and orange-outline nodes are sparse digitized targets such as boxes, body parameters, skeletons, keypoints, or audio tokens. Each line represents one task.

channels, never a full 256-channel tensor; all configured tasks currently use 115 channel IDs, so new types of world information can be added without changing the format. We used datasets including AVIS [40], Articulation-XL2.0 [41], ShapeNetSem [42], and MatSynth [43], and processed this data using our developed WSG dependency library, extracting information from RGB estimation and then injecting the RGB and additional information into WSG files.

These components contain at least the following information respectively: AVIS includes Depth and Spatial Geometry Domain [40] (relative depth [44] + absolute depth [45] + normal [46] + curvature + edge [47] + optical flow [48]), Pose and Human Information Domain [49–51] (dwpose + mediapipe + smpl + seg), and Sound And Language Properties Domain (Original Audio + Description Audio + Semantic Segmentation). Articulation-XL2.0 has Object/Animal Information Domain [52] (category + vertex + mesh + rig + uv + Joint). ShapeNetSem, in addition to similar basic obj information as the previous one, has Physics Properties Domain [53, 54] (relative solidVolume + relative surfaceVolume + relative weight + supportSurfaceArea + relative staticFrictionForce + material_id + material_confidence + density + staticFrictionCoeff + isContainerLike + Object category id) and Tradition and Control Optics of Material & Light Properties Domain (Material Name + Ambient light color + Diffuse reflection color + Opacity + Specular reflection color + hyperspectral index). MatSynth has PBR Material of Material & Light Properties Domain [43] (Basecolor + Diffuse + Normal + Height + Roughness + Metallic + Specular + Opacity). The processed datasets are available at WSM Datasets.

3.2. Four Generative Interfaces of Training

We organize WSG training as a directed graph centered at RGB. Text first enters visual space through $\text{text} \rightarrow \text{RGB}$. Around RGB, every WSG modality is a node, and every Channel Head[5] is a directed edge between nodes. As shown in Figure 2, the released dense WSM graph configures 3,042 visual task mappings around one shared generative backbone. The benchmark in this paper provides trained models and end-to-end inference results on the evaluation set for 1,148 dense Channel Heads; accordingly, 1,148 is the task count reported for WSM in the main paper. Sparse digitized Heads are evaluated separately in Table 12 and are not included in this count. Extending evaluation to the remaining dense mappings requires no change to the unified interfaces or model architecture.[6]

[5]The channel head is the final output component of our model. Due to the extremely rich modal information, we require specialized output heads for differentiation. The number of heads represents the number of tasks that the model can handle.

[6]If ResNet enables models to be trained especially deep, even with thousands of layers, our TunnelingNet enables models to be trained on

We use a pretrained text-and-image-to-image (TI2I) model as the starting point. Such models are commonly referred to as image-editing models and have already acquired the ability to take text and RGB as inputs and produce an image as output. The backbone is a DiT, so text, images, and conditions are all converted into tokens and concatenated into one input sequence. Based on our exploration, almost all visual tasks can be realized through four interfaces. An important question then arises: should RGB be sent to the native reference-image input for the model to modify it, or should it be used only as evidence/condition? We finally define the four interfaces as follows:

A RGB-to-World-Information estimation and reconstruction. This interface takes the currently observed RGB image as input; its target is World Information stored in channels additional to RGB, i.e., the non-RGB channels of WSG. This interface covers depth, normal, curvature, edge, optical flow, pose maps, semantic masks, material maps, physical properties, human structure, object structure, and visual-to-audio reconstruction. In the default dense graph, this interface uses only the current RGB image as condition: $\mathcal{F}_{\text{estimation}} : \mathbf{x}_{\text{RGB}}^{\text{current}} \rightarrow \mathbf{x}_{\text{world}}^{\text{target}}$.

B Text-and-World-Information narrow visual generation. Text supplies the semantic intent; the World Information condition supplies the structural control; and a reference RGB frame, placed in the input sequence, supplies the appearance control; during training, it is the first frame of the same video as the target frame. The corresponding non-RGB WSG channel enters a WSM condition lane, while the first RGB frame enters the native reference slot of the backbone. The target is the current RGB image, as in text-and-pose-to-video. In short, $\mathcal{F}_{\text{narrow generation}} : (\mathbf{y}^{\text{text}}, \mathbf{x}_{\text{world}}^{\text{condition}}, \mathbf{x}_{\text{RGB}}^{\text{first}}) \rightarrow \mathbf{x}_{\text{RGB}}^{\text{current}}$. And broad generation can be achieved through visual chains: $(\mathcal{F}_{\text{broad generation}} : \mathbf{y}^{\text{text}} \rightarrow \mathbf{x}_{\text{RGB}}^{\text{current}} \rightarrow \mathbf{x}_{\text{world}}^{\text{target}})$.

C World-Information-to-World-Information conditional control. Generated or observed World Information becomes the condition, and other World Information becomes the target (e.g., Edge to Normals): $\mathcal{F}_{\text{conditional control}} : \mathbf{x}_{\text{world}}^{\text{condition}} \rightarrow \mathbf{x}_{\text{world}}^{\text{target}}$.

D First-RGB-frame-plus-World-Information inverse rendering. A fixed first RGB frame gives the scene-level visual evidence, while observed World Information from the target time helps infer hidden World Information: shape, material, pose, audio source, or another world property. In the default dense graph, the RGB token 0–2 is routed through the first-frame condition branch, and `Reference_IMG` is not used (e.g., First RGB Frame as Condition and Normals to Edge). In short, $\mathcal{F}_{\text{inverse rendering}} : (\mathbf{x}_{\text{RGB}}^{\text{first}}, \mathbf{x}_{\text{world}}^{\text{observation}}) \rightarrow \mathbf{x}_{\text{world}}^{\text{hidden}}$.

Sparse Heads use the same input–output task definition as dense Heads, but need specialized output modules when the target is a list, box set, skeleton, body parameter, or audio token sequence rather than a dense image-like map. In this way, dense visual channels and sparse digitized structures are both treated as endpoints in the same generative WSG graph.

thousands of visual tasks.

Algorithm 1 Constructing and learning a dense modality-task edge

Require: edge $\mathcal{H}_k$, paired observations $\mathcal{D}_k$, shared generator f_θ
Ensure: task residual $\phi_k^\star$
$\rho_k \leftarrow \Gamma(\mathcal{H}_k)$ ▷ assign semantic roles
Associate boundary maps $\{\mathcal{E}_m : m \in \mathcal{C}_k \cup \mathcal{T}_k\}$
Couple task residual ϕ_k to the shared field f_θ
for $(\mathbf{x}_{\mathcal{C}_k}, \mathbf{x}_{\mathcal{T}_k}, \mathbf{p}) \sim \mathcal{D}_k$ **do**
  $\mathbf{z}_0 \leftarrow \mathcal{E}_{\mathcal{T}_k}(\mathbf{x}_{\mathcal{T}_k})$
  $(\mathbf{z}_{\mathcal{C}_k}, \mathbf{r}, \mathbf{p}) \leftarrow \Gamma_{\rho_k}(\mathbf{x}_{\mathcal{C}_k}, \mathbf{p})$
  Sample t and $\epsilon \sim \mathcal{N}(0, I)$
  $\mathbf{z}_t \leftarrow (1-t)\mathbf{z}_0 + t\epsilon$
  $\mathbf{v}^\star \leftarrow \epsilon - \mathbf{z}_0$
  $\widehat{\mathbf{v}} \leftarrow f_{\theta,\phi_k}(\mathbf{z}_t, t, \mathbf{z}_{\mathcal{C}_k}, \mathbf{r}, \mathbf{p}; \rho_k)$
  $\mathcal{L}_k \leftarrow \|\widehat{\mathbf{v}} - \mathbf{v}^\star\|_2^2$
  $\phi_k \leftarrow \phi_k - \eta \nabla_{\phi_k} \mathcal{L}_k$
end for
return $\phi_k^\star$

It is important to note that these four interfaces are not the Four Laws of Computer Vision. The four interfaces are the minimal common denominator of all tasks: visual tasks can be trained by using these interfaces as their training schemes. This does not imply an exact one-to-one correspondence with the four conceptual task categories described by the Four Laws. The Four Laws primarily divide the conceptual relationships between RGB and World Information into four broad categories, where B and B→A support narrow and broad Lex I; A supports Lex II; B and C support the RGB- and World-Information-target forms of Lex III; and A and D support reconstruction and inverse rendering under Lex IV. Image classification, localization, detection, segmentation, estimation, reconstruction, restoration, image and shape completion, image and point-cloud registration, pixel- and feature-level correspondence, tracking, retrieval, visual-database indexing, prediction, generation, editing, rendering, simulation, planning, and control all fall within these four categories.

3.3. Modality Architectures

A WSG mapping begins with a typed, directed relation that assigns every participating modality a semantic role:

$$\mathcal{H}_k = (\mathcal{C}_k, \mathcal{T}_k, \tau_k), \tag{1}$$

where $\mathcal{C}_k$ and $\mathcal{T}_k$ denote the condition and target modality sets and τ_k denotes one of the four generative interfaces. This relation induces a role assignment

$$\rho_k = \Gamma(\mathcal{H}_k) = (R_p, R_t, R_c, R_r, a_k), \tag{2}$$

which partitions information into prompt tokens R_p, a noised target stream R_t, clean modal evidence R_c, and an optional RGB appearance prior R_r; a_k selects the interaction rule between these roles. The construction therefore determines both the direction of prediction and the information available to it.

Each modality m has a boundary map $\mathcal{E}_m$ into the model's token space. Dense image-like fields and spatial surrogates share the normalized VAE boundary, while native sparse structures and audio use the compact adapters described below. After this boundary mapping, the learning principle is shared: the target follows a stochastic rectified-flow trajectory, whereas modal evidence remains clean. This clean-evidence/noised-target asymmetry turns a heterogeneous set of modalities into a directed generative problem. Algorithm 1 summarizes the construction for a dense task edge.

3.4. TunnelingNet

Around an RGB-based generative backbone, the WSG graph represents visual tasks as directed condition–target edges. A scalable WSM must add edges without replicating the full generator. TunnelingNet therefore factorizes learning into a shared World Snapshot generator and a collection of small Channel-Head tunnels. Given the typed edge $\mathcal{H}_k$ and its role assignment ρ_k, TunnelingNet models the transformation as

$$p_{\theta,\phi_k}\left(\mathbf{x}_{\mathcal{T}_k} \mid \mathbf{x}_{\mathcal{C}_k}, \mathbf{p}\right), \tag{3}$$

where θ is a frozen rectified-flow visual generator, ϕ_k is the trainable tunnel for Channel Head k, and $\mathbf{p}$ is the text prompt or modality prompt. The same backbone therefore supports many directed edges, and the marginal parameter cost of a new edge is the cost of its tunnel. WSM is one modular model comprising the shared backbone θ, a resident task-indexed bank of Channel Heads $\{\phi_k\}$, and an explicit task router. Heads can be optimized independently and physically sharded for parallel training and distribution; these are training and storage choices, not model boundaries. At inference, a task key sparsely activates one Head, while a visual chain activates a selected Head sequence. This yields one WSM with many generative paths.

A useful intuition is that TunnelingNet internalizes ControlNet: instead of attaching a large copied control branch outside the generator, it places low-rank residual steering inside every transformer block. MTA gives condition evidence its own independently normalized path rather than mixing it with target tokens in native attention. These layer-wise pushes redirect the full generation trajectory while leaving the backbone and evidence streams frozen, limiting disturbance to the native prior. Under matched trainable parameters, this routing reaches the selected structured outputs with 19.0–38.5× fewer updates and recognizable structure 77–100× earlier than LoRA NJA.

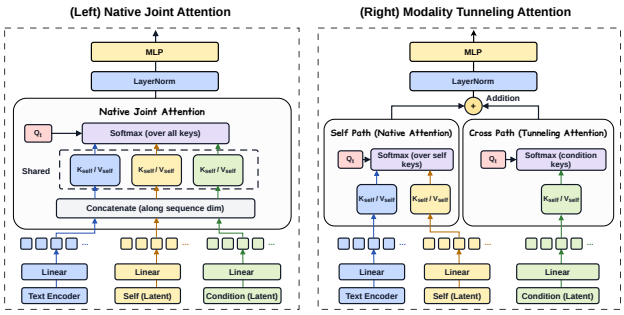


Figure 3. **Two attention routes in TunnelingNet.** (Left) **B** retains NJA, where text, target, condition, and RGB evidence share one joint attention field. (Right) Interfaces **A**, **C**, and **D** use MTA, which preserves target self-attention and adds an independently normalized route from conditions to the target.

As shown in Figure 3, a Channel Head separates what is learned from how evidence is routed. Its trainable component ϕ_k is a lightweight target-side tunnel, whereas its attention topology determines how clean evidence enters the noised target lane. For estimation/reconstruction, conditional control, and inverse rendering (Interfaces **A**, **C**, and **D**), we introduce *Modality Tunneling Attention* (MTA). MTA decomposes the target update into self-attention over prompt and target tokens and an independently normalized target-to-condition path:

$$\begin{aligned}\mathrm{Attn}_{\mathrm{tun}}(Q_t) = \mathrm{Attn}(Q_t, K_{\{p,t\}}, V_{\{p,t\}}) \\ + \mathrm{Attn}(Q_t, K_{\mathcal{C}_k}, V_{\mathcal{C}_k}).\end{aligned} \tag{4}$$

Here, $\mathcal{C}_k$ includes every active WSG condition and any RGB or reference observation routed as visual evidence. The evidence lanes themselves remain on the frozen native transformer path. MTA introduces no additional parameters; its separately normalized condition-to-target route allows a small target-side tunnel to learn heterogeneous non-RGB mappings efficiently.

We do not replace the native attention topology universally. For text-guided narrow visual generation (Interface **B**), whose target is RGB, we retain *Native Joint Attention* (NJA). NJA places text, target, condition, and the RGB evidence in one joint attention field, allowing RGB synthesis to reuse the backbone's image-generation prior.

In our implementation, θ is initialized from FLUX.2-klein-4B [31]. The text encoder, VAE, and base transformer weights are frozen. Each Channel Head is implemented as a small target-side tunnel placed on the transformer input embedding, modulation layers, attention projections, and feed-forward projections. Its trainable parameters are therefore

$$\phi_k = \{\Delta W^{(k)}_{\mathrm{embed}}, \Delta W^{(k)}_{\mathrm{mod}}, \Delta W^{(k)}_{\mathrm{attn}}, \Delta W^{(k)}_{\mathrm{ffn}}\}, \tag{5}$$

and optimization updates only ϕ_k:

$$\theta \leftarrow \theta, \qquad \phi_k \leftarrow \phi_k - \eta \nabla_{\phi_k} \mathcal{L}_k. \tag{6}$$

When the condition is an audio feature, the Channel Head also contains a small audio projection module that maps audio embeddings into the visual token space.

Training follows rectified flow in the latent space. Given a target latent $\mathbf{z}_0$ and noise ϵ, we sample t and construct

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\epsilon, \qquad \mathbf{v}^\star = \epsilon - \mathbf{z}_0. \tag{7}$$

TunnelingNet predicts the velocity from the noised target lane and clean condition lanes:

$$\mathcal{L}_k = \mathbb{E}\left[\left\| f_{\theta,\phi_k}(\mathbf{z}_t, t, \mathbf{z}_{\mathcal{C}_k}, \mathbf{r}, \mathbf{p}) - \mathbf{v}^\star \right\|_2^2\right]. \tag{8}$$

For deterministic dense targets, the implementation supports low-noise timestep shifting, Min-SNR weighting, Huber loss, high-noise anchoring, multi-resolution noise, and single-step clean-latent prediction. The default recipe uses high-noise anchoring for the four dense task types and keeps the other objective shaping switches off unless enabled. The task interface remains the same condition–target edge.

3.5. Deep Tunneling Learning

Why Deep Tunneling Learning Works. Figure 4 summarizes how Deep Tunneling Learning extends a standard DiT block while preserving its shared residual backbone. In Figure 4(a), the conditioning MLP produces the adaptive-LayerNorm scale and shift (γ_i, β_i) and residual gates α_i for the attention and pointwise feed-forward branches; the plus nodes denote ordinary residual addition. The two panels also retain the native objectives of their respective backbones: the standard latent-diffusion example predicts noise, while the FLUX.2 rectified-flow backbone predicts velocity. The $64 \times 64 \times 32$ tensors in panel (b) are FLUX.2 VAE latents for a 512×512 image. Latent shapes vary with the VAE rather than with the Tunneling Block; the standard-DiT example in panel (a) uses the smaller $32\times32\times4$ representation.

The method rests on the premise that heterogeneous visual modalities do not describe independent worlds. They expose different observations of the same underlying world state. A new modality mapping therefore need not reconstruct the visual generator; it needs a task-specific path that redirects the shared generative dynamics toward the target representation.

Why modalities differ even when they describe the same world. Consider two observers viewing the same object, one with typical color vision and the other with red–green color blindness. Their sensory representations differ, yet both remain grounded in the same object geometry, composition, and world state. Neither observer must reproduce the other's perception before using evidence supplied by the other. We express this shared-world hypothesis as

$$\mathbf{x}_m = g_m(\mathbf{s}), \qquad m \in \mathcal{M}, \tag{9}$$

where $\mathbf{s}$ is the latent world state and g_m is the observation map of modality m. WSG makes these heterogeneous observations addressable within one typed task graph, while the shared generator supplies their common visual prior.

How lightweight tunnels redirect generation. TunnelingNet preserves the representation learned by the shared generator and specializes its dynamics through a task-indexed residual field inside every transformer stage. For layer ℓ and module type $m \in \{\mathrm{embed}, \mathrm{mod}, \mathrm{attn}, \mathrm{ffn}\}$, the effective linear map of task k is

$$W_{\ell,m}^{(k)}\mathbf{h} = W_{\ell,m}\mathbf{h} + \frac{\alpha_k}{r_k} B_{\ell,m}^{(k)} A_{\ell,m}^{(k)} \mathbf{h}. \tag{10}$$

The shared term $W_{\ell,m}$ preserves the visual prior and a common coordinate system across modalities. The low-rank term changes only the vector field followed by the selected target modality. Under the shared-world hypothesis, a new task can therefore learn how to move through the common representation instead of learning that representation again.

How residual corrections accumulate across layers. For MTA, let $\mathbf{h}_\ell^t$ denote target tokens and $\mathbf{h}_\ell^{\mathcal{C}}$ denote the clean evidence tokens at block ℓ. Their interaction can be written as two residual components:

$$\mathbf{h}_{\ell+1}^t = \mathbf{h}_\ell^t + R_{\ell,k}^{\mathrm{self}}(\mathbf{h}_\ell^t, \mathbf{p}) + R_{\ell,k}^{\mathrm{evidence}}(\mathbf{h}_\ell^t, \mathbf{h}_\ell^{\mathcal{C}}). \tag{11}$$

The self residual advances the target along its generative trajectory, while the evidence residual corrects that trajectory using independently normalized observations. Repeating this correction through the transformer depth turns local updates into a complete modality tunnel: boundaries can constrain geometry, geometry can constrain surface orientation, and RGB evidence can constrain otherwise ambiguous world properties. Interface **B** retains the same task residuals but absorbs both interactions into NJA, preserving the native RGB-generation prior.

Shared backbone knowledge and task-specific learning. This construction separates two responsibilities. Frozen evidence pathways retain reusable observations in the shared feature space, while the selected target tunnel learns how those observations should alter one particular output modality. Deep Tunneling Learning consequently acquires new modality relations as sequences of small, task-indexed corrections while keeping the shared generator and previously learned tunnels unchanged.

3.6. Categorical Output Head

For semantic categorical targets, a Channel Head includes an integrated categorical output path. The rectified-flow

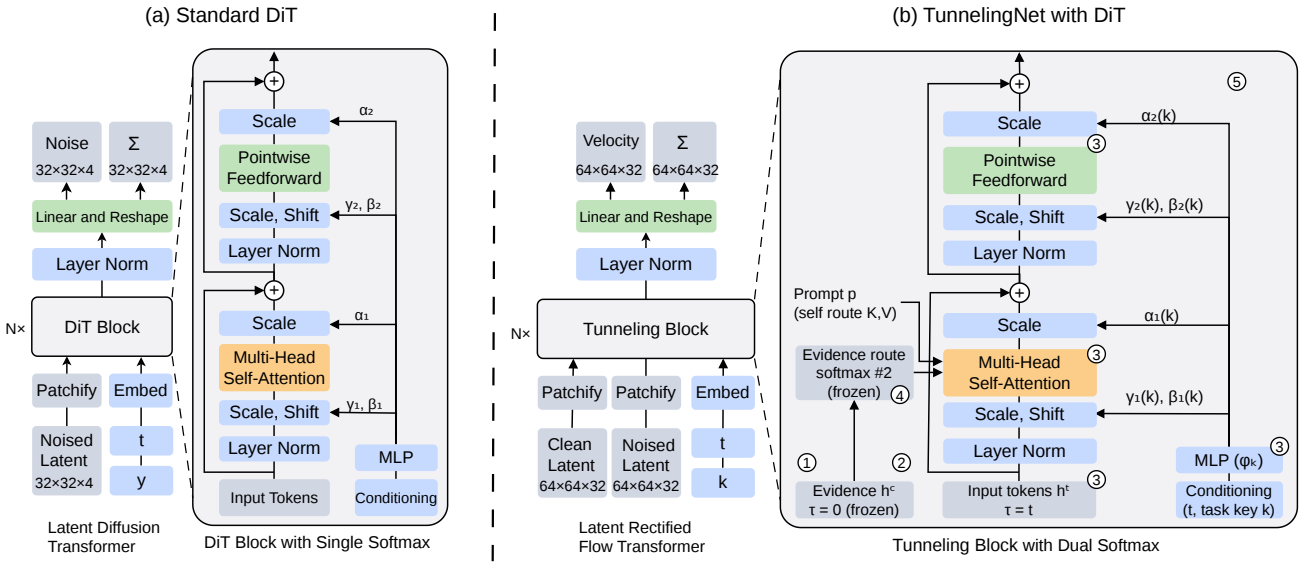


Figure 4. **From a standard DiT block to the Tunneling Block.** Panel **(a)** shows the standard DiT block, and panel **(b)** its task-indexed conversion. ① Target, evidence, and prompt tokens receive distinct computational roles. ② The target follows flow time $\tau = t$, whereas clean evidence remains fixed at $\tau = 0$. ③ The task key k selects Channel Head ϕ_k, whose task-specific updates cover the input embedding, conditioning MLP, attention projections, and pointwise feed-forward projections. ④ In MTA, target tokens provide queries Q_t; prompt and target tokens provide keys and values $K_{\{p,t\}}, V_{\{p,t\}}$ for the self route, while evidence provides keys and values K_{c_k}, V_{c_k} for a second, independently normalized Softmax route. ⑤ The same pattern repeats through all N Tunneling Blocks. Together, these changes reuse and extend a generative backbone into heterogeneous modality spaces with almost zero inter-task interference.

branch predicts spatial support, while a lightweight classification branch assigns the exact semantic palette value to each generated region. These branches define one deployable Head; the intermediate pseudo-color field is an internal representation rather than a separate model output. Continuous fields and edge maps use only the flow path.

Let $\hat{\mathbf{x}}_k$ be the spatial field produced inside categorical Channel Head k, and let $\{R_j\}$ denote its foreground connected components. For each region, we mask the encoded condition used by the task:

$$\mathbf{h}_j = \mathrm{Concat}(\mathbf{z}_{\mathcal{C}_k}) \odot \mathrm{Resize}(R_j). \tag{12}$$

A compact spatial encoder pools $\mathbf{h}_j$ into a region embedding, which is compared with learned category prototypes using cosine similarity:

$$\ell_{j,c} = \exp(s) \left\langle \overline{e_\psi(\mathbf{h}_j)}, \overline{\mathbf{p}_c} \right\rangle. \tag{13}$$

During training, regions and labels are obtained directly from the ground-truth categorical map. We use class-balanced cross-entropy and select the classifier state with the highest validation balanced accuracy. At inference, regions are extracted from the generated spatial field, classified independently, and written with the exact palette value of the predicted category. This naturally supports multiple objects and categories in one image.

The complete categorical prediction therefore preserves the generative localization while resolving category identity from the task condition. Its classification parameters are stored with the corresponding Channel-Head run and loaded automatically by evaluation and inference. For a categorical target, both parts are required; a checkpoint without its classification parameters is not treated as a deployable categorical Head.

3.7. Sparse-Structure Modeling

We find that dense and sparse prediction need not form two independent task systems. Their difference can instead be confined to the representation boundary of a Channel Head. When a sparse structure admits a spatial surrogate, we encode it as an image-like field and retain the same latent-flow learning path used by dense WSG modalities. For example, keypoint tuples are converted to semantic Gaussian heatmaps, while instance boxes are represented as filled pseudo-color regions. After generation, peak extraction, non-maximum suppression, or connected-component analysis recovers the original sparse structure.

Some outputs, such as body parameters, object skeleton trees, and audio tokens, do not have a natural dense representation. These tasks use compact structure-specific adapters instead. Dense visual evidence is converted only at the adapter boundary—for example into full-image features, person-crop features, or masked modality features—

and the predicted structure is written back into its designated WSG channels. Consequently, a dense WSG prediction can condition a sparse task, and a reconstructed sparse result can re-enter the same world representation without introducing a separate task ontology.

The shared generative backbone is initialized from FLUX.2-klein-4B, while its transformer, text encoder, and VAE remain frozen. Dense tasks update only their target-side tunnels using AdamW in bfloat16. Training and visualization share a portable set of normalization rules, which are inverted before values are written back to WSG. Categorical output branches are trained for 3,000 updates, validated every 50 updates, and retain only the checkpoint with the best balanced accuracy.

Table 2. **Final optimal configuration selected by our ablation evaluation.** MTA and NJA denote Modality Tunneling Attention and Native Joint Attention, respectively; RGB_c and RGB_1 denote current- and first-frame RGB. All four heads train only the target tunnel, freeze the evidence lanes, and use high-noise anchoring without additional dense-objective shaping.

Head	Interface	Condition → target	Attention	RGB evidence route
003	Estimation/reconstruction	$\text{RGB}_c \rightarrow \text{Normal}_c$	MTA	RGB_c condition
270	Narrow generation	Text + Normal_c + $\text{RGB}_1 \rightarrow \text{RGB}_c$	NJA	Native reference
275	Conditional control	$\text{Normal}_c \rightarrow \text{Edge}_c$	MTA	None
313	Inverse rendering	RGB_1 + $\text{Normal}_c \rightarrow \text{Edge}_c$	MTA	RGB_1 condition

4. Ablation Evaluation

We organize the experimental study into three main sections. This section establishes the final TunnelingNet design through controlled ablations; Section 5 measures large-scale task coverage and efficiency; and Section 6 examines model outputs across the four generative interfaces.

Our ablation evaluation has three primary objectives: **(I) Attention Topology**, determining how target and condition tokens should interact; **(II) Evidence Routing and Tunnel Placement**, establishing how condition/reference evidence should enter the model and which tunnels require adaptation; and **(III) Rectified-Flow Objective Shaping**, identifying an objective formulation that remains effective across heterogeneous output modalities. Together, these objectives determine the architecture, evidence pathways, and optimization recipe of the final TunnelingNet.

To ensure that the methods selected for these three objectives are broadly effective across task types rather than specialized to a single mapping, we choose four representative Channel Heads from our broader task suite, with one task from each interface: RGB-to-Normal estimation (Head 003), text-and-Normal-to-RGB narrow generation (Head 270), Normal-to-Edge conditional control (Head 275), and first-RGB-plus-Normal-to-Edge inverse rendering (Head 313). Table 2 summarizes these four representative heads and the final optimal configuration identified by our three-part ablation evaluation. Together, these tasks span dense geometric and photorealistic RGB outputs, sparse structural outputs, and distinct uses of current observations and first-frame visual evidence. We apply each ablation to the representative heads for which the corresponding design choice is applicable. Agreement across interfaces supports a generally effective choice, whereas consistent differences identify where interface-specific treatment is necessary. Unless stated otherwise, the FLUX.2-Klein-4B backbone, text encoder, and VAE are frozen; each head activates only 12.5M target-side tunnel parameters, or approximately 0.32% of the 3.9B-parameter visual backbone. We use the same data split and fixed validation set within each controlled sweep. Loss values across different target modalities are reported to establish convergence, not to rank the intrinsic difficulty of the modalities.

4.1. Attention Topology

In this ablation, NJA uses the same target-side low-rank parameterization as MTA and serves as our LoRA native-attention baseline. MTA changes only the attention routing; the trainable parameter count, data, optimizer, batch size, and training schedule remain fixed. Native Joint Attention performs one softmax over the pooled text, target, condition, and reference keys. In a controlled Head-270 comparison at 140 updates, NJA obtained a validation MSE of 0.1917, compared with 0.2155 for MTA, an 11.0% reduction. We therefore retain NJA for narrow visual generation. Figure 5(a) pairs the directly recorded validation curves with the condition, selected NJA output, and target at update 100. This result reflects the different requirements for conditioning information in RGB generation and other modality estimation tasks. Modality Tunneling Attention aims to limit the influence of the input RGB on the target generation path, allowing conditioning information to act on the target through an independent one-way channel. This mechanism is suitable for estimation, simple conditional control, and inverse-rendering tasks because it can preserve the target modality's own generation process and prevent the input evidence from dominating. However, conditional RGB generation typically requires edges, normals, or other conditions to make larger changes to the structure and content of the reference RGB, and therefore requires fuller interaction among the condition, text, reference image, and target sequences. Native Joint Attention places these sequences in the same attention space, allowing every token to influence the entire generation sequence and thereby enabling the conditioning information to globally reorganize the RGB generation path. Therefore, Native Joint Attention is better suited to conditional RGB generation, whereas Modality Tunneling Attention is better suited to non-RGB estimation and reconstruction tasks that require the scope of the input influence to be controlled.

The other three interfaces have heterogeneous, often deterministic non-RGB targets. Here, joint normalization forces target self-attention and condition attention to compete for the same probability mass. We identify this competition as the central optimization bottleneck and introduce MTA to remove it. Rather than requiring heterogeneous modalities to be jointly understood within one shared atten-

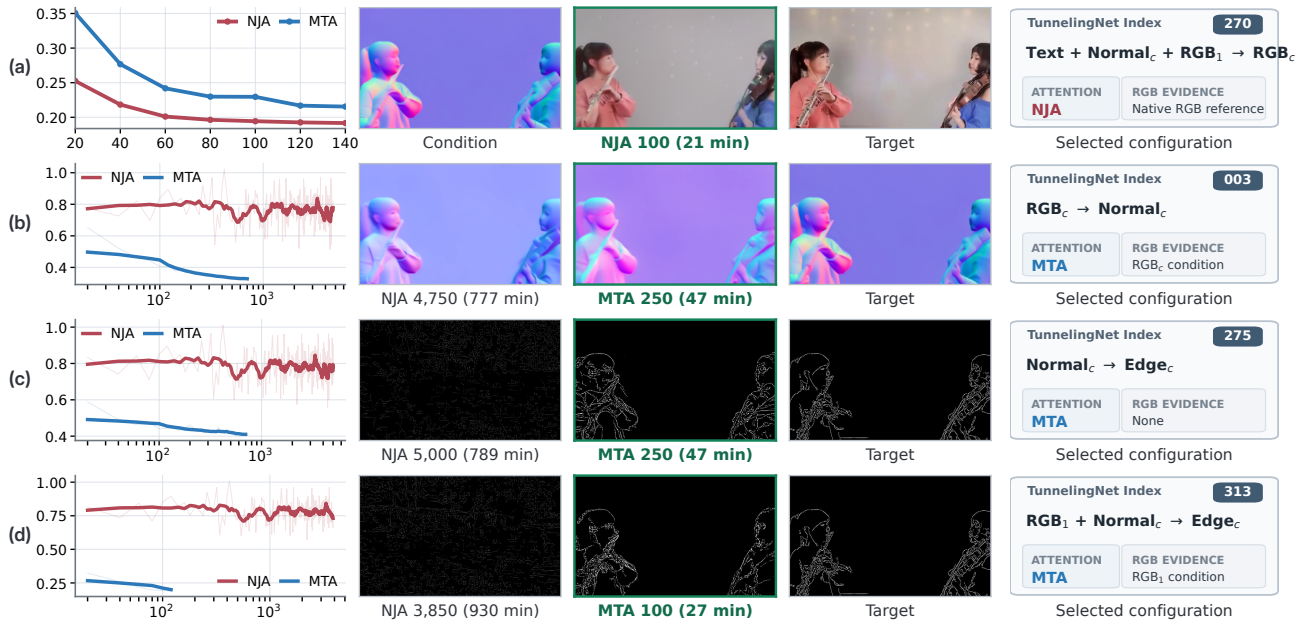


Figure 5. **Attention ablation and selected TunnelingNet configurations.** Rows show (a) narrow generation, (b) estimation/reconstruction, (c) conditional control, and (d) inverse rendering. Curves plot validation MSE (lower is better) against optimizer updates; thin/thick traces denote measurements/nine-point moving averages, and (b–d) use logarithmic horizontal axes. Under matched parameters, data, optimizer, batch size, and schedule, NJA is preferred for the RGB target. For the three structured non-RGB targets, MTA produces clean, recognizable outputs in 27–47 minutes with 19.0–38.5× fewer updates, while NJA remains pale or fragmented after 777–930 minutes; recognizable structure appears 77–100× earlier. These endpoint time gaps of 16.6–34.1× are conservative because NJA has not reached comparable quality. The legacy NJA previews in (c,d) use different evaluation examples; all Edge maps use the same grayscale reader. The rightmost cards summarize the deployed mapping, attention topology, and RGB-evidence route.

tion distribution, our method retains each modality's own representation path and learns a separately normalized tunnel between different modal descriptions of the same underlying visual content. Concretely, MTA preserves the text–target self path and adds an independently normalized target-to-condition path.

The effect is most visible for sparse, high-frequency Edge prediction. With batch size one and no gradient accumulation, MTA recovers recognizable Edge structure after roughly 50 updates, corresponding to only 50 sampled training examples, and produces clean boundaries by update 250. NJA also learns the mapping, but its structural acquisition remains at a rudimentary stage even after thousands of updates. RGB-to-Normal exhibits the same contrast: NJA converges much more slowly and produces pale, fog-like normals with weak boundaries. Figure 5(b) shows that MTA obtains a structured Normal prediction by update 250, while NJA requires 4,750 updates. Figures 5(c,d) show the same effect on conditional control and inverse rendering: MTA recovers clean Edge structure by updates 250 and 100, whereas NJA has only begun to produce a few fragmented short contours after 5,000 and 3,850 updates, respectively, remaining at an early stage of learning. Thus, the LoRA NJA baseline has sufficient capacity to learn these mappings; its limitation is optimization efficiency. Parameter efficiency alone does not remove the competition between target and condition evidence inside native joint attention.

The modest, non-monotonic NJA validation MSE is expected: this global velocity metric can favor mean or background solutions and therefore need not track decoded structural quality. The curves and samples must be read together. Measured by optimizer updates, MTA reaches the displayed structured outputs at least 19.0×, 20.0×, and 38.5× sooner than the same-parameter LoRA NJA baseline on Heads 003, 275, and 313, respectively. The first recognizable structure emerges approximately 77–100× earlier. Because the NJA Edge outputs remain lower quality at their cited endpoints, the corresponding update ratios are conservative lower bounds. MTA introduces no additional parameters. These results also show why one attention topology should not be imposed on all four interfaces: MTA is preferable for transferring structured non-RGB evidence, while NJA is preferable when synthesizing RGB with the native image prior.

Table 3. RGB-routing ablations for estimation and narrow generation. (a) Head 003 follows the final 200-update recipe except for the current-RGB route. (b) Head 270 uses the same legacy protocol for all variants and is evaluated at update 2,220. Comparisons are controlled within each block; absolute MSE should not be compared across blocks. *Target* indicates a trainable target-side tunnel. Bold marks the lowest MSE, while † marks the retained route.

(a) Head 003: RGB_c →$Normal_c$

Attention	RGB_c route	Target	Val. MSE ↓
MTA	Condition branch†	Yes	**0.1844**
MTA	Native reference	Yes	0.1853

(b) Head 270: Text + $Normal_c$ + RGB_1 →RGB_c

Attention	RGB_1 route	Target	Val. MSE ↓
NJA	Native reference†	Yes	**0.6277**
NJA	Condition branch	Yes	0.6279
NJA	Native reference	No	0.6311
MTA	Condition branch	Yes	0.6481

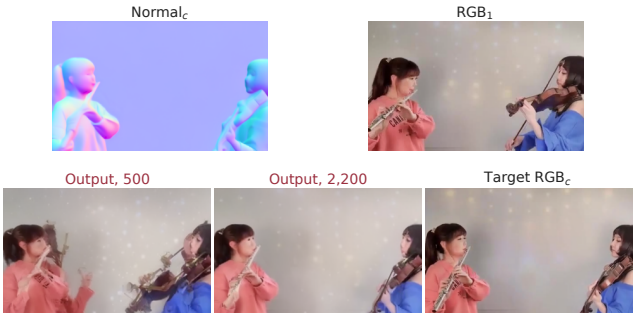


Figure 6. **Native-reference routing can override current geometry.** True 10-step inference from the Head-270 native-reference run. At update 500, conflicting poses produce duplicate limbs; at update 2,200, the output still retains the RGB_1 pose instead of following the current Normal.

4.2. Condition Routing and Tunnel Placement

Routing is nontrivial because the pretrained backbone introduced in Section 3 is an image-editing model. Its native interface accepts a text prompt and an optional RGB reference—the image to be edited—and uses the resulting preservation prior while generating the target. TunnelingNet additionally exposes clean condition lanes for WSG modalities and RGB observations. The same RGB evidence can therefore enter either as a native editable reference or as a condition: the former encourages the backbone to preserve the supplied image, whereas the latter treats it as evidence rather than content to copy. A second choice is tunnel placement: we can adapt both the condition and target lanes, or train only the target-side tunnel while leaving all evidence lanes on the frozen backbone.

These choices yield a different candidate set for each representative task. **Head 003 (estimation)** maps RGB_c to $Normal_c$; we test whether the current RGB should occupy the native editable-reference slot or a current-RGB condition lane. **Head 270 (narrow generation)** synthesizes RGB_c from text, $Normal_c$, and RGB_1; we test whether the first frame should be a native appearance reference or a condition, and whether a target-side tunnel is needed. **Head 275 (conditional control)** maps $Normal_c$ to $Edge_c$ without any RGB input, so its evidence route is fixed and it isolates whether the Normal condition lane itself needs a trainable tunnel. **Head 313 (inverse rendering)** predicts $Edge_c$ from current $Normal_c$ and RGB_1; we test whether RGB_1 should use the native reference slot or a separate first-frame condition lane, and whether adapting the condition side improves over a target-only tunnel.

Together, these alternatives define our routing-and-placement sweep. Table 2 consolidates the final configuration for all four heads, Table 3 reports the controlled RGB-routing comparisons for Heads 003 and 270, and Figure 7 summarizes the corresponding qualitative choices, including tunnel placement for Head 313.

The Head-003 block of Table 3 compares the two current-RGB routes for Head 003 under the same 200-update MTA recipe. The condition branch obtains 0.1844 MSE, compared with 0.1853 for the native reference. The small 0.43% gap shows that both routes can transmit RGB evidence, but the condition route is marginally better and better matches the estimation task: the current RGB provides the visual evidence from which the Normal map is inferred, rather than an appearance reference that the image-editing backbone is encouraged to preserve.

The Head-270 block summarizes the controlled routing sweep. The two NJA variants with trainable target tunnels are numerically tied: routing RGB_1 through the native reference slot improves MSE by only 0.0002 (0.03%). Figure 6 reveals the more important distinction: the native slot strongly preserves the reference appearance and can resist changes specified by the current Normal, as if the task were direct image editing. Head 270 generates the current RGB from text, current Normal, and first-frame appearance evidence. We retain NJA for its RGB prior and select the native reference route, which obtains the slightly lower validation MSE and matches the final configuration of the selected Head. Removing the target tunnel weakens the direct adaptation path to RGB, while replacing NJA with MTA gives up part of the native RGB synthesis prior. Figure 7(b) fixes Normal, RGB_1, and the target across both routes. The condition-branch result is true 10-step inference at update 200 (45 min 32 s) and illustrates that the alternative route can follow a deliberately conflicting pose while retaining first-frame appearance.

Inverse rendering makes this separation stricter. In Head 313, current Normal defines the target state, whereas RGB_1 is only a scene-level clue. Routing RGB_1 as a native reference encourages a shortcut that copies stale appearance

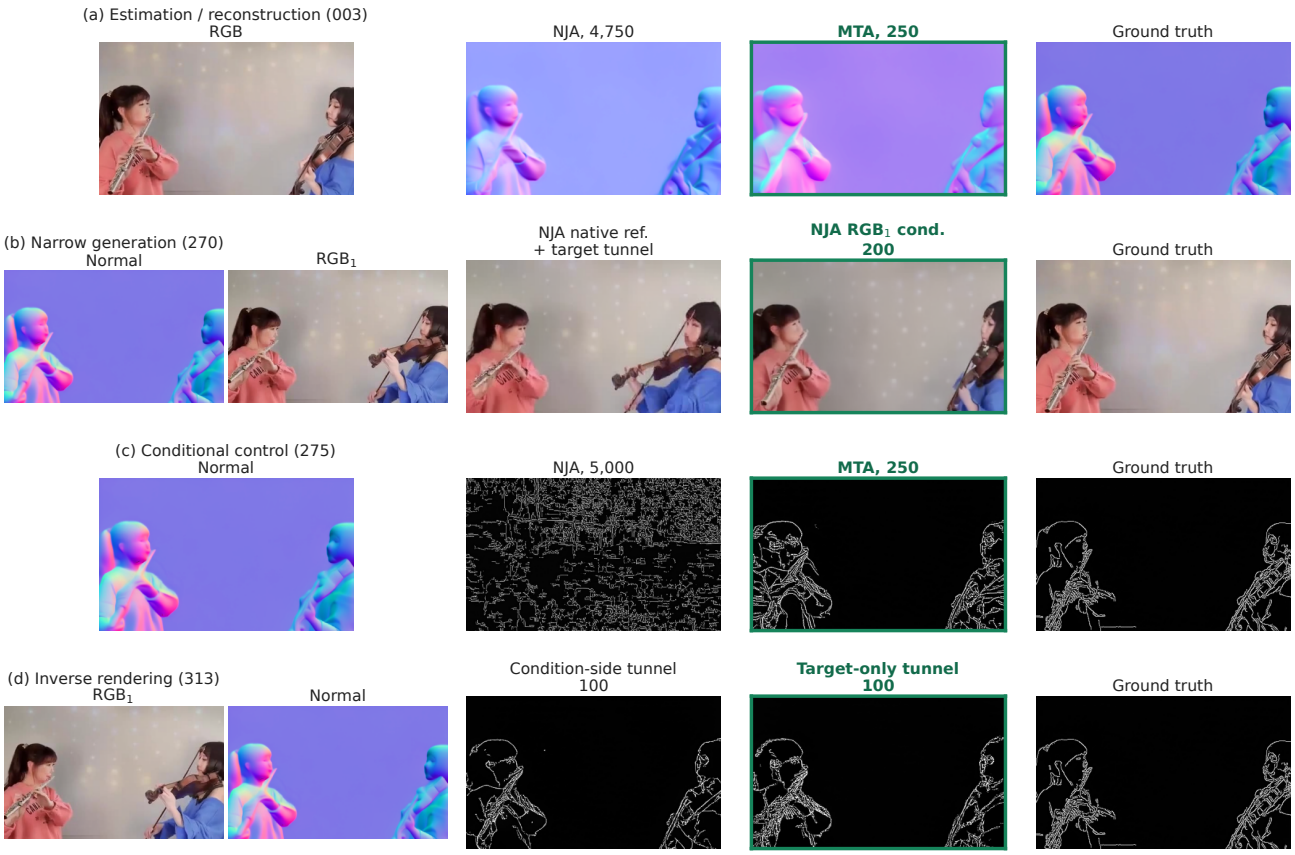


Figure 7. **Qualitative effects of task-dependent structural choices.** Each row uses an evaluation example from one of the four interfaces. (a,c) compare NJA with MTA for estimation/reconstruction and conditional control. (b) compares native-reference and first-RGB-condition routing under NJA for narrow generation. (d) compares condition-side and target-only tunnel placement for inverse rendering. Green borders mark the visually preferred outputs in the displayed comparisons. The highlighted outputs in rows (b) and (d) use 200 and 100 optimizer updates, reached in 45 min 32 s and 27 min 18 s, respectively. In (b), the condition-branch result follows the Normal/Pose condition while retaining the appearance supplied by RGB_1; the deployed Head follows the slightly lower validation MSE and uses the native reference route reported in Table 2. Checkpoint indices are shown above outputs.

Table 4. Task-agnostic objective ablation. Values are fixed-$t = 1$ validation velocity MSE, so all rows are directly comparable.

Objective	003 Normal	275 Edge	Mean ↓
Plain rectified flow	0.6805	0.5961	0.6383
Single-step clean latent	0.4308	0.4186	0.4247
High-noise anchor	**0.3263**	**0.4062**	**0.3662**
Multi-resolution noise	0.6505	0.5917	0.6211

or geometry. We instead pass its tokens through the first-frame condition branch and disable the native reference slot, so the current Normal remains the state-defining evidence. Accordingly, Head 270 is the only selected configuration that uses the native RGB-reference slot. Heads 003 and 313 route RGB as condition evidence, while Head 275 has no RGB input.

Figure 7(d) isolates tunnel placement under the same 100-update budget. Both variants recover comparable major Edge structure, and adapting the condition side adds no visible structural gain. This does not mean that condition evidence is ignored: clean Normal and RGB_1 tokens still enter the separately normalized MTA path. Rather, the result locates the head-specific learning burden at the target readout, which must translate already informative frozen evidence into the requested output modality. It is the practical counterpart of the color-vision analogy in Section 1: modality-specific streams need not acquire a common internal interpretation, or learn to "understand one another," to describe the same underlying state. The target tunnel learns only the translation needed by its output. Removing condition-side reinterpretation therefore reduces the optimization burden and keeps condition states adapter-independent, allowing one condition forward pass to be shared by many Channel Heads. We consequently train only the target tunnel, the smallest effective design. The panel also reveals the complementary role of visual

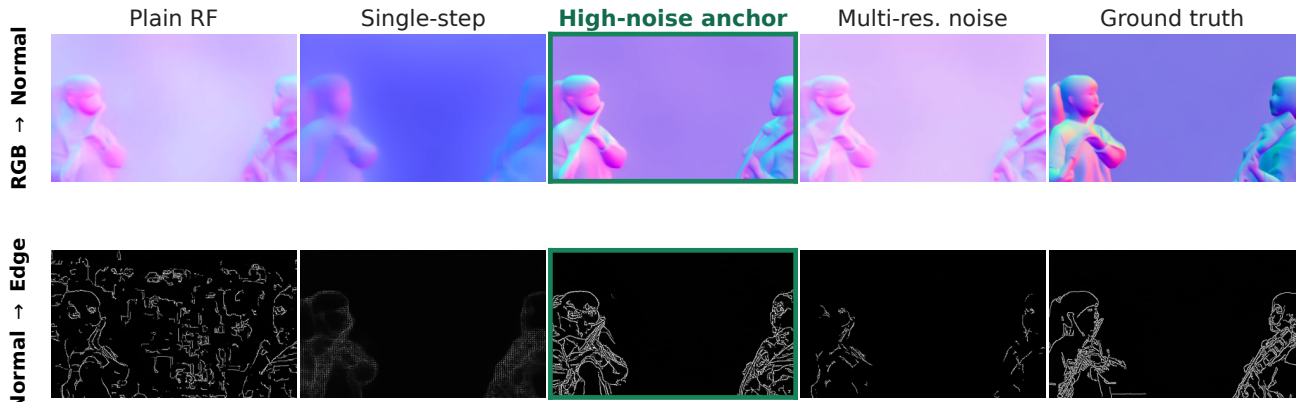


Figure 8. **Qualitative effects of rectified-flow objective shaping.** At update 250, each row compares plain rectified flow, single-step clean-latent prediction, high-noise anchoring, multi-resolution noise, and the ground truth. With batch size one and no gradient accumulation, each model has completed only 250 optimizer steps and trained on exactly *250 sampled examples*. High-noise anchoring is the only shared choice that preserves both Normal saturation and sparse Edge structure. Green borders mark the selected objective; all Edge maps use the same grayscale reader.

evidence: the Normal estimate alone omits fine structures smoothed away during estimation, whereas providing RGB_1 as additional evidence enables the inverse-rendering head to recover them. This cross-modal completion provides an empirical basis for Chain of Visual (CoV) reasoning, in which evidence retained in one visual state resolves information absent from another. With routing and tunnel placement fixed, the remaining objective-shaping choice is quantified in Table 4 and analyzed next.

4.3. Rectified-Flow Objective

The initial dense-task experiments exposed two failure modes. Normal predictions retained the correct coarse shape but drifted toward low-saturation mean colors; sparse Edge predictions were progressively overwhelmed by their black background. We therefore tested task-agnostic objective changes rather than modality-specific losses. Table 4 compares plain rectified flow, single-step clean-latent prediction, high-noise anchoring, and annealed multi-resolution noise. Every model is evaluated by the same plain velocity MSE at $t = 1$. High-noise anchoring, which assigns half of the training samples to $t = 0.999$, is best on both representative target types and reduces their mean validation error by 42.6% relative to the baseline. Figure 8 shows that this advantage is not merely numerical. For RGB-to-Normal, plain rectified flow and multi-resolution noise recover the coarse layout but weaken the chromatic separation that encodes surface orientation; the anchored model retains a more saturated, spatially coherent field. For Normal-to-Edge, the same alternatives lose much of the sparse contour support to the dominant black background, whereas anchoring preserves connected structure. Single-step prediction is competitive in Table 4, but its decoded outputs are less consistent across these dense and sparse targets. High-noise anchoring is thus the only tested objective that improves the shared validation metric while also preserving the modality-specific information visible after decoding.

We also ablated low-noise timestep shifting, Min-SNR weighting, and Huber loss. Huber produced the cleanest early visual results, and Min-SNR better resisted late degradation, but neither was a consistent default. In particular, enabling all three gradually drove outputs toward the target mean: Normal kept its shape but faded, while Edge disappeared into the black background. This qualitative behavior is not captured reliably by a weighted training loss, which is why we retain fixed plain-MSE validation and inspect the decoded output throughout training. The final concise recipe is therefore high-noise anchoring alone, with timestep shifting, Min-SNR, Huber, single-step prediction, and multi-resolution noise disabled. With this recipe, the four representative heads reach stable solutions within a 500-update ablation budget, compared with the 5,000-update budget used by the earlier native setup.

4.4. Minute-Scale Rapid Learning

Figure 9 tracks end-to-end inference for all four interfaces at 10-update intervals through update 50 and again at update 200. Because batch size is one, update count is also the number of sampled training examples. The first predictions are obtained after only 1 min 19 s to 3 min 35 s, the 50-update milestone after 19 min 32 s to 47 min 35 s, and the 200-update endpoint after 44 min 27 s to 2 h 13 min 27 s. This wall-clock spread reflects the different conditioning contracts and their diagnostic overhead; update count provides the cleaner measure of data exposure, which is identical across rows.

Using a fixed evaluation example and inference seed

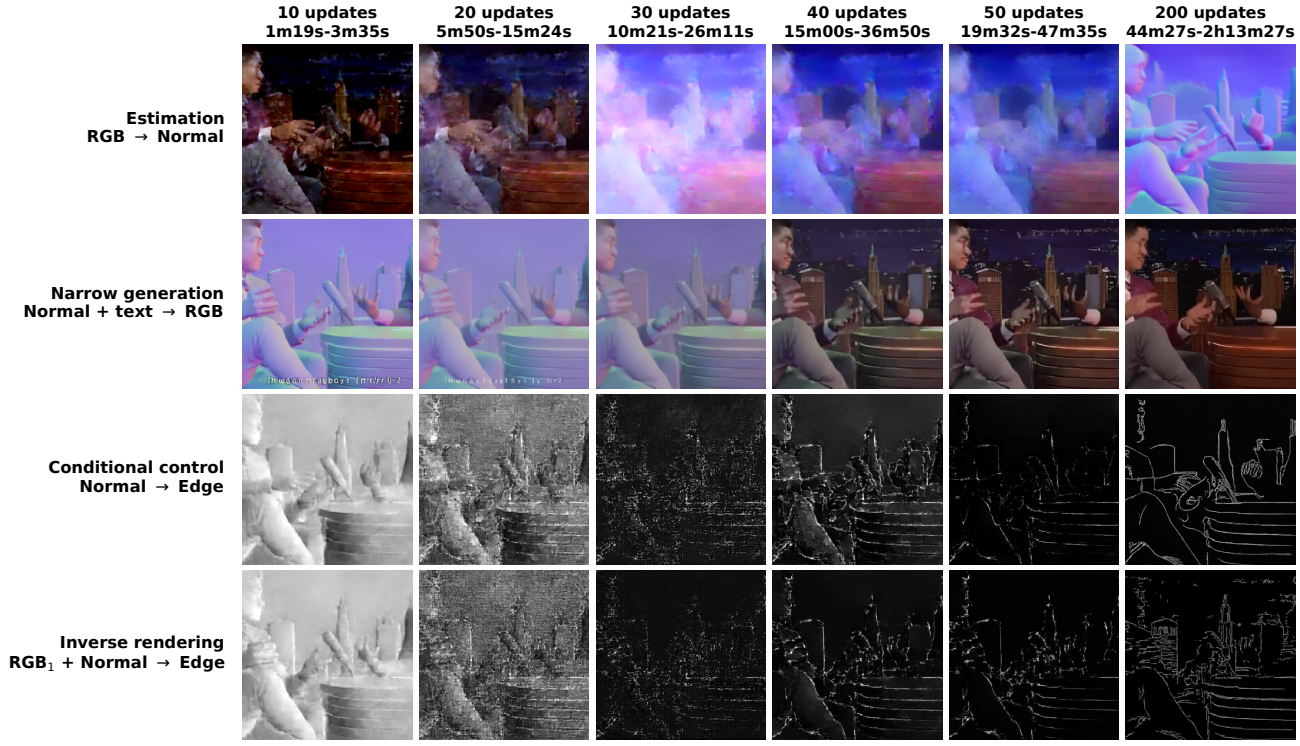


Figure 9. **Rapid acquisition across the first 200 updates.** Rows correspond to estimation, narrow generation, conditional control, and inverse rendering. Columns show true 10-step Euler inference from pure noise after 10, 20, 30, 40, 50, and 200 updates. Times below the column headings report the observed wall-clock range across the four tasks, including scheduled diagnostics. Each row reuses one fixed evaluation example and the same inference seed. With batch size one and no gradient accumulation, *each update uses exactly one sampled training example*; thus even the final column has seen only 200 examples. No condition or ground-truth image is displayed. Edge outputs use the same per-image contrast stretch for visibility.

within each row, the sequence reveals a consistent coarse-to-structured transition. The first tens of updates establish the target modality and broad spatial support; later updates sharpen task-specific geometry, improve structural continuity, and suppress residual artifacts. This trend appears across dense estimation, RGB synthesis, sparse conditional control, and inverse rendering despite their different output statistics. Most importantly, the update-200 predictions have been trained on only 200 sampled examples. They are therefore evidence that the proposed routing and objective acquire a new visual mapping at minute scale, rather than requiring conventional long-horizon fine-tuning. These early outputs are not treated as final convergence; Appendix Figure 50 provides condition–prediction–target comparisons at the later selected checkpoints.

Table 5. **Coverage of the frozen dense evaluation snapshot.** Input signatures count distinct conditioning contracts, not individual images. All outputs are obtained by true 10-step Euler inference.

Interface	Heads	Inputs	Targets	Evaluations
RGB estimation / reconstruction	38	1	38	692
Text + WSG narrow generation	38	38	1	704
WSG conditional control	536	37	37	8,411
RGB-assisted inverse rendering	536	37	37	8,384
Total	1,148	–	39	18,191

5. Large-Scale Quantitative Evaluation

Protocol. We freeze one deployable checkpoint for each trained dense Channel Head and evaluate it end to end from pure Gaussian noise through the complete generation path. Every run uses 10 Euler steps, a fixed head-specific seed, and deterministically selected examples from the evaluation set. Across 1,148 Heads, the resulting snapshot contains 18,191 end-to-end inference evaluations. Table 5 reports coverage across the four interfaces.

Metrics and task matrices. For continuous or image-like channels we report decoded pseudo-RGB MSE and SSIM. Because modality-specific encodings have different numeric scales, MSE is compared only among Heads sharing the same output channel. For semantic categorical outputs, the main matrices report dominant-class accuracy (DAcc) from the final deployed Head–refiner route. DAcc measures whether the principal generated region receives the correct identity. The released tables additionally retain mIoU as a stricter spatial and multi-region metric; strict pixel accuracy remains secondary because background can dominate it. The complete input–output matrices retain every trained edge numerically. Figure 10 first presents the 76 RGB-centered Heads by pairing the two directions for each of 38 WSG modalities: RGB-to-WSG estimation in the upper-left and text+WSG-to-RGB narrow generation in the lower-right. The remaining 1,072 Heads form 536 shared input–output pairs across conditional control and RGB-assisted inverse rendering. Figures 11– 14 place both results in each diagonally divided cell rather than repeating the axes: conditional control occupies the upper-left and RGB-assisted inverse rendering the lower-right.

Continuous results use the higher-is-better quality score QS $= 1 -$ decoded MSE, where decoded MSE is computed between the prediction and ground truth in the normalized VAE pseudo-RGB decoder space; categorical cells show dominant-class accuracy (DAcc). Triangles are colored by QS for continuous outputs and DAcc for categorical outputs. Released tables retain exact metrics for all mappings, including those below the display threshold. To avoid a mostly empty global rectangle, we partition the trained support into five compact blocks: RGB-centered

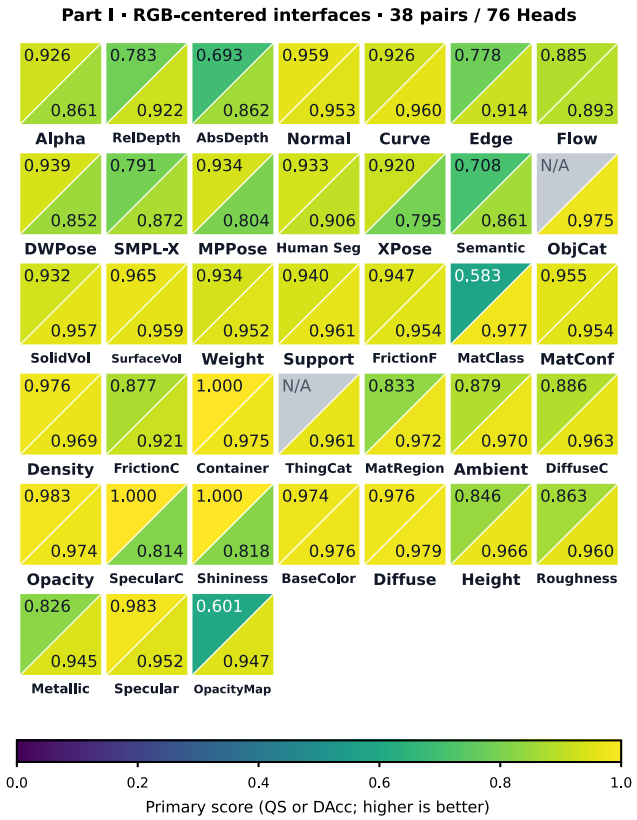


Figure 10. **Part I: RGB-centered interfaces (38 modality pairs; 76 Heads).** Each cell pairs upper-left RGB-to-WSG estimation/reconstruction with lower-right text+WSG+RGB$_1$-to-RGB narrow generation. Continuous entries use QS and categorical estimation entries display DAcc. Their triangle color uses DAcc; continuous triangle color uses QS on the same zero-to-one scale used throughout Figures 10– 14. Cells marked N/A are unavailable in the current release.

interfaces, physical and material properties, visual structure and pose, PBR texture space, and cross-family bridges. These Parts define the layout—not a task taxonomy or method components—and summarize 1,034 Heads. Mappings unavailable in the current release are marked N/A.

5.1. Cross-Modal Matrix Analysis

Performance by task direction. Figure 10 exposes a consistent target-side asymmetry. RGB estimation reaches a median primary score of 0.886, whereas narrow RGB generation reaches 0.953; 68.4% and 97.4% of the respective Heads exceed 0.8. The representation explains this gap: RGB synthesis can reuse the pretrained appearance prior, while RGB-to-WSG translation must reproduce the exact geometry or pseudo-color convention of each target channel. The same tendency recurs in the paired matrices, where the RGB-assisted direction is usually stronger than condition-only control. It therefore measures the value of

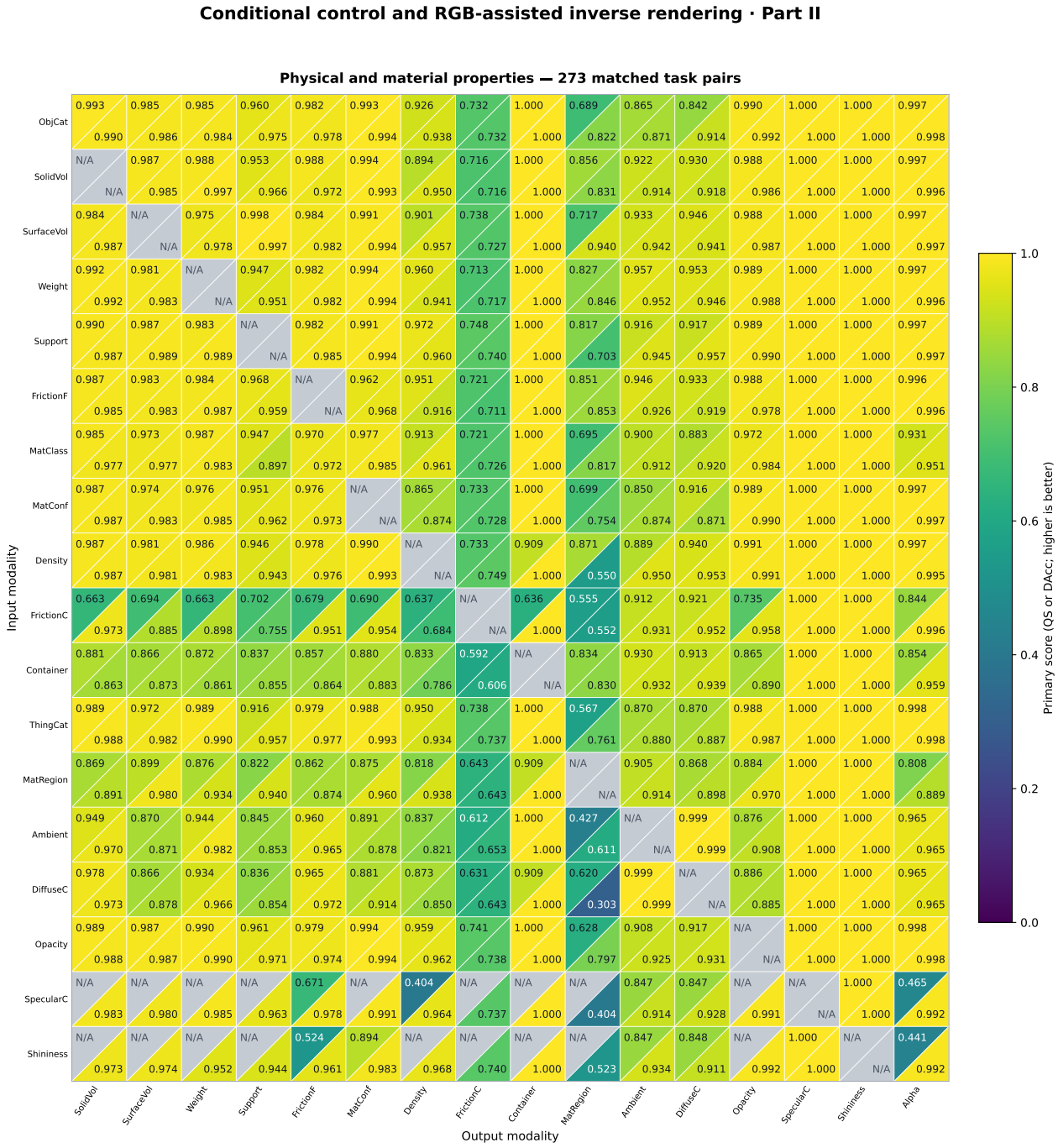


Figure 11. **Paired cross-modal matrix, Part II: physical and material properties (273 pairs; 546 Heads).** Each cell places conditional control in the upper-left triangle and RGB-assisted inverse rendering in the lower-right. Continuous entries use QS $= 1 -$ decoded MSE; categorical entries display DAcc. Triangle colors encode QS for continuous outputs and DAcc for categorical outputs on a common zero-to-one scale. Cells marked N/A are unavailable in the current release.

a visual reference, not an intrinsic ranking of the two interfaces.

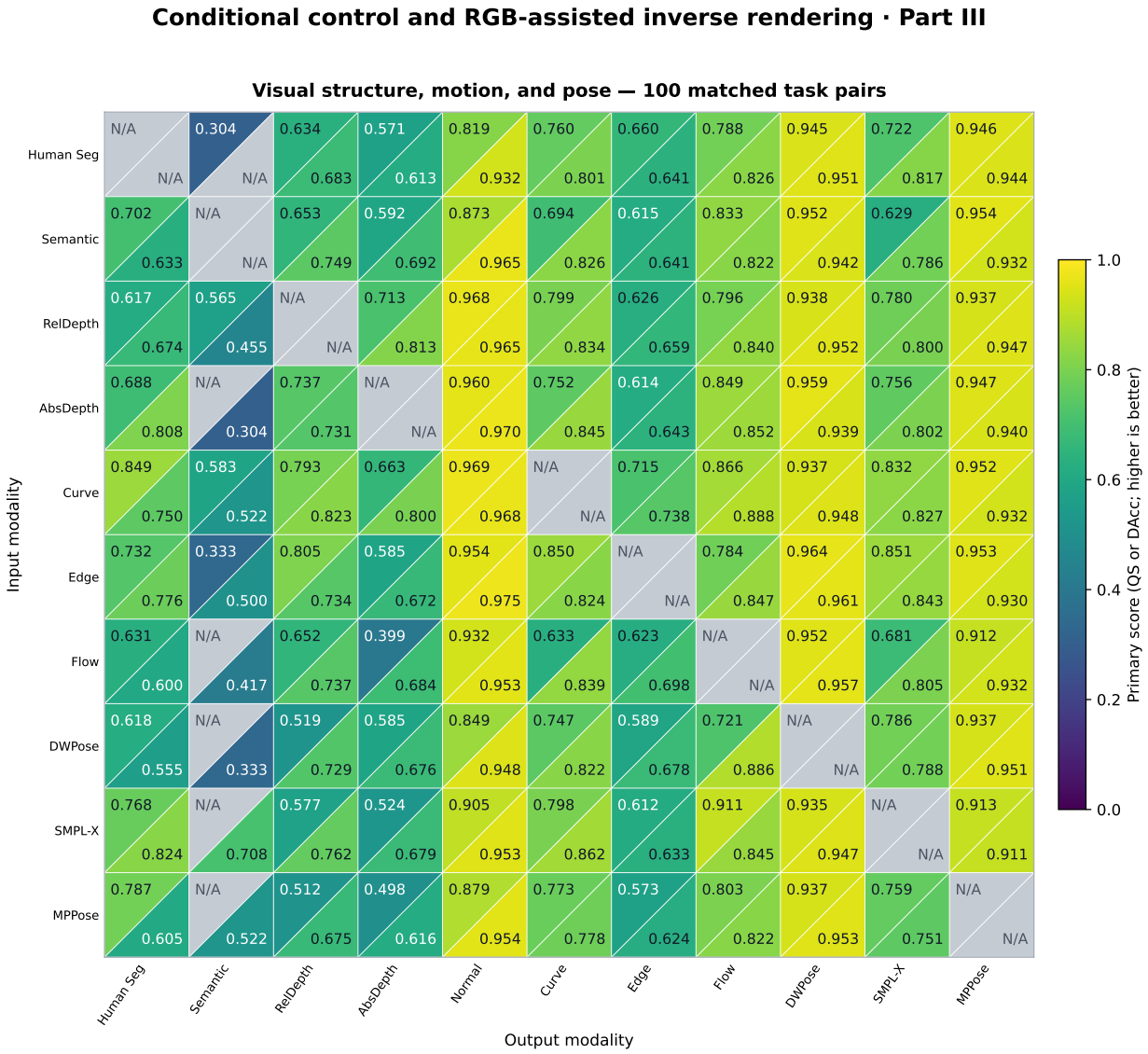


Figure 12. **Paired cross-modal matrix, Part III: visual structure, motion, and pose (100 pairs; 200 Heads).** Diagonal cell ordering and metrics follow Figure 11. Separating this family from the global matrix enlarges pose, geometry, and semantic annotations while preserving their original input–output coordinates. Cells marked N/A are unavailable in the current release.

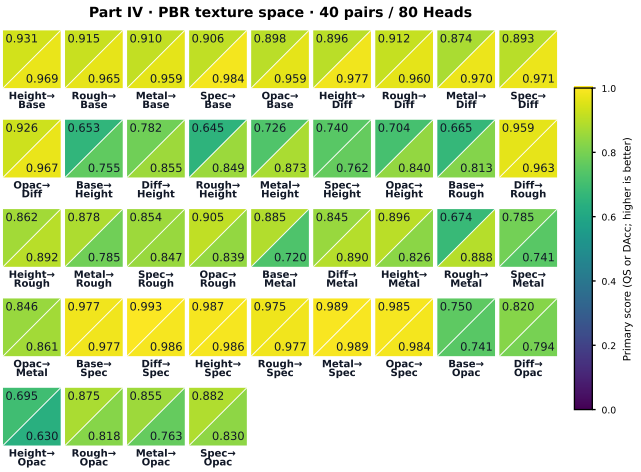


Figure 13. **Part IV: PBR texture space (40 pairs; 80 Heads).** The compact card layout retains only trained input–output pairs; upper-left and lower-right triangles report conditional control and RGB-assisted inverse rendering. Cells marked N/A are unavailable in the current release.

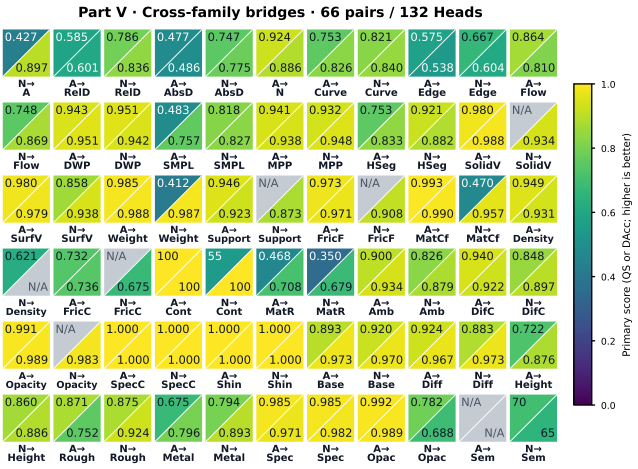


Figure 14. **Part V: cross-family bridges (66 pairs; 132 Heads).** The companion compact layout wraps the trained Alpha- and Normal-input pairs. Categorical cells show percentage DAcc and use DAcc for color; continuous cells show and color by QS. Cells marked N/A are unavailable in the current release.

Physical and visual structure. The physical/material block in Figure 11 is the strongest large family: its median score is 0.947 for conditional control and 0.973 for RGB-assisted inverse rendering, with 79.5% and 89.0% of mappings above 0.8. Spatially smooth scalar fields and appearance channels dominate this high-score region. Friction coefficient, material region, and the categorical container mask are harder, which separates approximate spatial support from exact semantic calibration. Figure 12 shows a more selective pattern. Its medians are 0.798 and 0.839, but Normal and articulated pose outputs remain consistently strong: Normal reaches median scores of 0.919 and 0.960, while DWPose reaches 0.945 and 0.951. This pattern is consistent with human geometric reasoning: Depth and Edge cannot be recovered reliably when the conditioning modality does not preserve the corresponding spatial cues.

PBR attributes and cross-modal conversion. Figure 13 is nearly directionally symmetric. The median scores are 0.880 for control and 0.881 for inverse rendering, and 72.5%/77.5% of its mappings exceed 0.8. Base color, diffuse, and specular channels transfer particularly reliably, while height and opacity expose the remaining direction-dependent cases. The bridge matrix in Figure 14 clarifies when a reference image matters. Alpha-conditioned mappings remain strong in both directions (median 0.932/0.933), whereas Normal-conditioned mappings rise from 0.790 under condition-only control to 0.897 with RGB assistance. The RGB reference thus supplies appearance and scene-specific evidence that a structural channel alone does not contain, while the structural condition still determines the requested transformation.

Interpretation. Taken together, the matrices support a structured claim: a single tunneling mechanism covers hundreds of heterogeneous transfers, and performance follows the information shared by the input and target modalities. Large coherent regions indicate reusable cross-modal transport; directional gaps identify mappings that benefit from first-frame RGB evidence; and N/A cells keep unavailable mappings from being mistaken for valid low-confidence predictions. We retain their exact records in the released tables; an extended analysis of categorical limitations and the corresponding upgrade inventory is provided in Appendix 10.1.

5.2. Prediction–Target Correspondence Calibration

The matrix scores above compare each generated output with its corresponding held-out target. We additionally measure how strongly those scores depend on the correct sample correspondence. For every Head, we keep its predictions fixed and cyclically shift the targets by a deterministic nonzero offset chosen away from adjacent samples. This produces a mismatched control with the same prediction set, target set, modality, and decoder statistics. We then compute MSE for the matched and shifted pairings in the normalized decoder space used by QS. The calibration requires no retraining or additional inference.

Table 6. **Prediction–target correspondence across all four interfaces.** Error reduction compares the mean matched and shifted MSE.

Interface	Error reduction
RGB estimation / reconstruction	57.2%
Text + WSG narrow generation	78.3%
WSG conditional control	33.5%
RGB-assisted inverse rendering	46.3%
All interfaces	**41.4%**

Correct target correspondence lowers mean error in every interface. Across the full matrix, it reduces error by 41.4%, and the median per-Head reduction is 51.9%. A fixed output template is unchanged when targets are permuted and would therefore produce zero correspondence reduction. The strictly positive reduction in every interface, including a greater-than-half median reduction across Heads, directly rejects that behavior. The per-Head and per-target records and deterministic calibration script accompany the evaluation tables. These consistent gaps show that the matrix scores capture sample-conditioned target structure beyond modality-level averages or fixed output templates. The calibration therefore establishes nontrivial prediction–target correspondence across all four interfaces of the large task graph.

Table 7. **Three complementary levels of evidence for the Unified Law.** The evidence asks whether the law works broadly, whether the trained routes learn the requested relation for each example, and why one target representation is easier for a frozen generator to learn than another. Together, they establish broad applicability, genuine task learning, and the mechanism that makes aligned representations easier to acquire.

Level	The question	What we did	What it shows
I: broad use	Was the law fitted only to the four tasks that first revealed it?	We established the four interfaces with four initial Tunnels, then kept the shared backbone frozen and trained the other 1,144 task routes.	The same rule realizes 1,148 trained and evaluated tasks, covering 39 output groups and all four interfaces.
II: real input–output learning	Did each route learn the requested relation for each example, or only a typical output for its modality?	We keep every prediction fixed and deliberately pair it with the wrong held-out target. A generic output would be equally close before and after this shift.	Correct pairs have 41.4% lower error overall, with a 51.9% median reduction per Head. The routes therefore follow the individual input and target.
III: why representation matters	Does a frozen generator learn the same information more easily when it is written in a representation the generator already knows?	We write the same surface-normal target in five lossless coordinate systems: its native VAE coordinates and four reversible channel rearrangements. Every version reconstructs exactly the same normal image; only its familiarity to the backbone changes.	Native coordinates lower checkpoint-averaged error by 20.1%–26.6%, produce the lowest final error, and win 59 of 64 final sample–transformation comparisons.

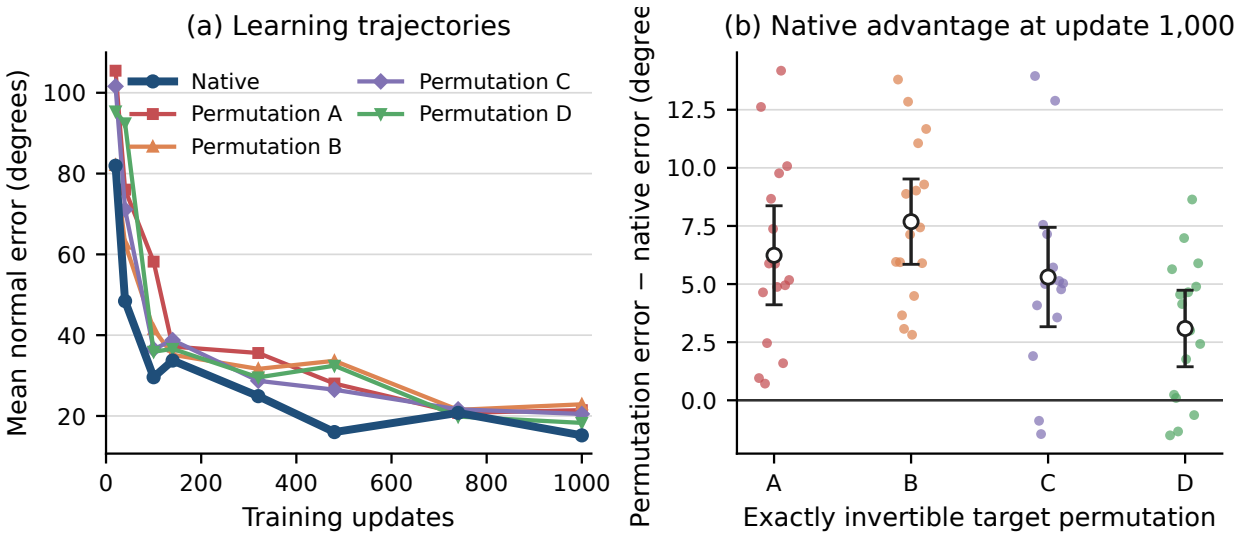


Figure 15. **The same target is easier to learn in coordinates already familiar to the frozen generator.** We encode every surface-normal target in its native VAE coordinates and in four reversible rearrangements of the same channels. All five versions retain exactly the same information and reconstruct the same normal image. The rearrangements simulate alternative representations that the backbone did not learn during pretraining. **(a)** Native coordinates give the lowest checkpoint-averaged error across the training trajectory and the lowest final error. **(b)** Positive values mean that native coordinates have lower final angular error for one held-out sample under one rearrangement. Native coordinates improve all four mean comparisons and 59 of the 64 individual comparisons. This directly shows that backbone familiarity changes learning difficulty even when target information is identical.

5.3. Three Complementary Levels of Evidence: Breadth, Genuine Task Learning, and Mechanism

What this section establishes. The same rule extends from four initial tasks to 1,148 trained tasks; the resulting routes learn the requested relation for individual examples; and a target becomes substantially easier to learn when it uses a representation already understood by the frozen generator. Together, these results support the Unified Law as both a way of organizing visual tasks and a learning principle. The law explains the large-scale success of WSM, predicts which representation of a new target will be easier to acquire, and provides a practical rule for building future task routes.

This subsection asks three progressively harder questions. First, did the rule continue to work after the four tasks that revealed it? Second, did the resulting routes learn the requested relation for each example? Third, can we directly test the proposed reason that some representations are easier for a frozen generator to acquire? The three levels move from breadth, to genuine input–output correspondence, to a controlled test of the learning mechanism. Table 7 gives the complete logic in one place.

Level I: the rule extends far beyond the tasks that revealed it. We first developed Heads 003, 270, 275, and 313, one representative task for each of the four interfaces. These four experiments established how a task should be expressed and where a TunnelingNet should connect to the shared generator. We then applied that rule to the other 1,144 routes. Each route was trained and evaluated while the shared generator remained frozen. The result is 1,148 trained tasks across 39 output groups. This order matters: thousands of results were obtained after the common rule had already been formed, so the rule is not a summary invented separately for each finished task. The complete chronological record, including sparse, categorical, and pose variants, appears in Sec. 14.12.

Level II: the routes learn each requested relation, not only the appearance of a modality. A model could produce a plausible depth map, normal map, or mask without following the particular input image. We test this possibility without changing any prediction. For every Head, we first compare each prediction with its true held-out target. We then shift the targets between examples and compare the same predictions with targets belonging to other inputs. An output that only repeats a generic modality pattern receives essentially the same error after the shift. A prediction that follows its own input loses agreement. Section 5.2 shows the latter: correct pairing reduces error in every interface, by 41.4% over the whole task graph, and the median Head loses 51.9% of its agreement after targets are shifted. The thousand-task result therefore reflects sample-specific visual relations.

Level III: the same information is easier to learn when its representation matches the backbone. The first two levels show that the law works broadly and that the routes learn real relations. The third level tests why representation choice changes the difficulty of learning. We use RGB-to-surface-normal estimation and write every normal target in five different coordinate systems. The first is the native VAE representation seen by the generator during pretraining. The other four simply rearrange the 32 VAE channels. Each rearrangement is perfectly reversible: it removes no information and reconstructs exactly the same normal image after the inverse rearrangement. It therefore simulates a modality encoded in a complete but unfamiliar organization.

This comparison separates two ideas that are often confused. A representation can encode and decode a target perfectly, while a pretrained backbone may still be unfamiliar with how that representation organizes the information. Here all five versions are equally complete; only their match to the frozen generator changes. We keep the task, data, backbone, TunnelingNet, optimization, update budget, and normal evaluator identical. Thus, any consistent learning advantage comes from representation–backbone alignment.

Before training the five Tunnels, we ask the frozen backbone to predict each coordinate system. Its residual is lowest for the native coordinates (1.249, versus 1.653–1.835), confirming that these are the coordinates most familiar to the pretrained generator. We then train all five versions and measure the same native normal-angle error at eight shared checkpoints on 16 held-out examples. The result follows the prediction. Native coordinates have the lowest checkpoint-averaged error (23.39°, versus 29.28°–31.85°), a 20.1%–26.6% reduction. They also achieve the lowest final error (15.20°, versus 18.29°–22.89°) and win 59 of 64 final sample–transformation comparisons. Figure 15 shows both the full trajectories and the per-example results.

Takeaway: the law predicts both what can be unified and how to make it learn efficiently. The three levels establish one connected result. A rule discovered from four Tunnels extends to 1,148 trained visual tasks. Those routes learn the requested relation for individual examples across all four interfaces. When the target information is held exactly constant, the representation already understood by the frozen generator is learned substantially faster and more accurately. The last result provides controlled evidence for the comparative prediction in Eq. 20.

This gives the Unified Law practical force. Encoding and decoding a modality is only the first requirement. Efficient capability expansion also places the target in a representa-

tion whose organization is familiar to the shared backbone. Dense visual targets can often use the backbone's native visual coordinates directly. Sparse or highly abstract targets benefit from a visual reorganization or a task-appropriate output boundary. This principle explains the successful choices observed across the development record and provides a concrete way to design new Tunnels before paying for full training. The formal statement of this result, its information-preservation and backbone-coverage conditions, and its extension to other generative backbones appear in Sec. 7.5.

Table 8. **External comparison on representative tasks across three WSM interfaces.** Evaluation uses the shared AVIS evaluation set. External methods run released official checkpoints, and each WSM request activates its registered task route. Each quality metric occupies its own column. Parameters report the active route; WSM activates the 4B base and one 12.5M-parameter TunnelingNet. Training Sample Budget is the number of optimizer updates multiplied by the global batch size, including all devices and gradient accumulation. Relative Sample Budget is the ratio of processed examples to the corresponding WSM Head. Frame-based models retain their native training units. Training Time/Hardware gives the training cost reported in each paper or official repository. WSM uses the elapsed time to the selected checkpoint on one A100, including periodic validation. Hardware-specific values remain in their original units. For MODUS, the reported token budgets, per-device packed sequence lengths, and 64-GPU launch configuration recover approximately 68.8K optimizer updates and 4.40M packed sequences; because a packed sequence may contain multiple examples, its reported ratio is a lower bound. Scaling Diffusion Perception reports 4K updates after MoE upcycling, but does not disclose the preceding mixed-task update count, global batch size, or wall time. N/R denotes a factor not reported by the paper or recoverable from the official code. Bold marks the best reported value in each arrow-directed column within a task. Inference Steps counts denoising/sampling updates (GeoWizard reports 50×10 for 50 steps and an ensemble of 10; feed-forward VGGT uses one pass). Total Inference Time is summed steady-state inference time over the shared evaluation set; Inference Speed is recomputed from the same timing. Baseline timing excludes model loading and uses consecutive prediction writes. WSM timing includes amortized loading and output decoding. All measured single-GPU runs use an RTX 8000; SenseNova FP32 timing uses four RTX 8000 GPUs. Only award designations are colored red. † marks methods that additionally receive the same first-frame RGB reference.

A Estimation and reconstruction: RGB→Depth. Native normalized depth in $[0, 1]$ is used; external predictions receive a fixed 2nd–98th-percentile linear map, depth direction is aligned before scoring, and WSM uses its direct output.

Method	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	MAE↓	RMSE↓	PSNR↑	Acc.05 ↑	Acc.10 ↑	Acc.20 ↑	Pearson↑	Spearman↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
Marigold [55]	CVPR 2024 Best Paper Award Candidate	0.9B	18K×32	331.0	A few days 1 GPU	0.247	0.291	11.22	0.183	0.293	0.441	0.890	0.875	1×10	105.47	0.2181
DMP [56]	CVPR 2024	0.86B+LoRA	50K×8	229.9	N/R	0.277	0.324	10.14	0.119	0.211	0.386	0.736	0.693	5	24.38	0.9434
GeoWizard [57]	ECCV 2024	0.9B	20K×256	2.94K	Time N/R 8 GPUs	0.308	0.384	8.72	0.237	0.345	0.440	0.818	0.795	50×10	7834.00	0.0029
Scaling Diffusion Perception [16]	CVPR 2025	2.8B (1.0B active)	Mixed-task phase N/R +4K MoE updates	N/R	N/R	0.232	0.274	12.22	0.235	0.344	0.508	0.868	0.825	50×3	833.92	0.0276
VGGT [58]	CVPR 2025 Best Paper Award	1.2B	160K×3,072 frames	N/R	9 days/64 A100s 13.8K A100-h	0.243	0.296	10.98	0.193	0.297	0.453	0.719	0.698	**1 pass**	**4.83**	**4.7619**
MODUS [15]	ICML 2026	14B (7B active)	∼ 68.8K×64 packed seq.	≥ 2.53K	88.5 h/64 GH200s 5,664 GH200-h	0.167	0.216	14.04	0.335	0.479	0.666	0.867	0.799	50	9025.88	0.0025
SenseNova-Vision [59]	Technical Report	7B	50K steps (packed)	N/R	Time N/R ≥ 16×80GB GPUs	0.171	0.217	13.90	0.314	0.444	0.637	0.856	0.834	50	3471.00 (4G)	0.0066
WSM	Ours	4B+12.5M	1740 × 1	**1.0**	11.0 h 1 A100	**0.130**	**0.162**	**17.22**	**0.415**	**0.590**	**0.770**	**0.935**	**0.907**	10	222.41	0.1034

B Conditional RGB generation: Edge→RGB. Canny thresholds are $100/200$.

Method	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Edge Density Error ↓	MAE↓	MSE↓	PSNR↑	SSIM↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
ControlNet [11]	ICCV 2023 Best Paper Award	1.4B	N/R×256 (v1.1)	N/R	3 days/8 A100s (v1.1)	0.0286	0.276	0.120	9.90	0.127	30	**79.17**	**0.3032**
IP-Adapter [17] (with ControlNet [11])†	Technical Report	2.0B	ControlNet: N/R×256 IP-Adapter: 1M×64	≥ 145.5K	≥ 1,176 A100-h total ControlNet: 3 days/8 A100s IP-Adapter: time N/R/8 V100s	0.0387	0.242	0.100	10.72	0.271	30	87.21	0.2752
T2I-Adapter [60]	AAAI 2024	2.6B+79M	20K×256	11.6K	Time N/R 4 A100s	0.0265	0.281	0.118	9.84	0.060	30	262.46	0.0914
UniControl [61]	NeurIPS 2023	1.5B	∼ 900K×64	∼ 130.9K	> 5,000 A100-h A100-40G	0.0298	0.293	0.129	9.36	0.058	50	265.64	0.0903
Uni-ControlNet [62]	NeurIPS 2023	1.5B	1 epoch (10M pairs)	22.7K	N/R	0.0597	0.284	0.118	9.77	0.083	50	381.89	0.0628
WSM†	Ours	4B+12.5M	440 × 1	**1.0**	8.5 min 1 A100	**0.0157**	**0.087**	**0.021**	**19.51**	**0.762**	**10**	202.56	0.1185

C Any-to-any control over representative structured signals. Edge, relative depth, surface normals, and human-part semantics span geometric boundaries, spatial layout, local surface shape, and structured human understanding.

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Accuracy↑	FPR↓	Specificity↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Normal→Edge	ICML 2026	14B (7B active)	∼ 68.8K×64 packed seq.	≥ 2.20K	88.5 h/64 GH200s 5,664 GH200-h	0.8671	0.0569	0.9431	50	12345.09	0.0019
WSM	Normal→Edge	Ours	4B+12.5M	2000 × 1	**1.0**	9.3 h 1 A100	**0.8797**	**0.0296**	**0.9704**	**10**	**239.76**	**0.1001**

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Mean Angle (°) ↓	Median Angle (°) ↓	11.25° ↑	22.5° ↑	30° ↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Relative Depth →Normal	ICML 2026	14B (7B active)	∼ 68.8K×64 packed seq.	≥ 2.25K	88.5 h/64 GH200s 5,664 GH200-h	35.19	31.13	0.222	0.431	0.512	50	12272.18	0.0020
WSM	Relative Depth →Normal	Ours	4B+12.5M	1960 × 1	**1.0**	9.6 h 1 A100	**26.40**	**22.48**	**0.231**	**0.527**	**0.664**	**10**	**1767.4**	**0.0136**

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Pixel Accuracy↑	mIoU↑	Human IoU ↑	mAcc↑	Macro-F1↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Edge→Human Segmentation	ICML 2026	14B (7B active)	∼ 68.8K×64 packed seq.	≥ 2.32K	88.5 h/64 GH200s 5,664 GH200-h	0.8431	0.6577	0.5148	0.7676	0.7630	50	9100.77	0.0026
WSM	Edge→Human Segmentation	Ours	4B+12.5M	1900 × 1	**1.0**	9.6 h 1 A100	**0.8862**	**0.6918**	**0.5391**	**0.7753**	**0.7780**	**10**	**1579.62**	**0.0152**

5.4. External Comparison on the AVIS Evaluation Set

Comparison scope. Table 8 covers five representative established design paradigms: new trainable parameter blocks (ControlNet), the MoE-style multi-expert adapter in UniControl, LoRA (DMP), expert-based MoE-family architectures—the mixture-of-transformers in MODUS and strict sparse-MoE diffusion in Scaling Diffusion Perception—and IC-Wrapper methods (T2I-Adapter and Uni-ControlNet). WSM introduces internal trajectory tunneling. The three blocks correspond to Interfaces A–C: RGB-to-depth estimation, Edge-to-RGB generation, and any-to-any control. We also include VGGT as a task-specific depth estimator. Interface C evaluates transfers among edge, relative depth, surface normal, and human-part segmentation signals. Both WSM and MODUS support a broader transfer inventory; we select these tasks as a representative cross-section of structured understanding, spanning object and region boundaries, scene-scale spatial organization, local surface orientation, and human semantics. **Across all three externally comparable interfaces and all five representative tasks, WSM leads every reported quality metric, demonstrating that its advantage holds across all five adaptation paradigms.** Interface D combines first-frame RGB evidence with a later auxiliary observation, for which we found no released image-level method with matching inputs. Most remaining mappings enabled by this interface similarly define new task settings without a like-for-like external comparator.

Evaluation data. Images from widely used benchmark datasets may have been included in the web-scale pretraining corpora of large foundation models. We therefore use the AVIS evaluation set, which is less commonly used for dense prediction, and generate the task annotations evaluated here ourselves. This choice reduces the likelihood of direct overlap with existing image–annotation pairs. None of the evaluation images or annotations is used to train WSM. A semantic analysis of the scene captions also finds almost no close match between the evaluation scenes and the WSM training descriptions. Every method receives the same AVIS inputs and targets. The external methods use their released checkpoints, and WSM uses its trained Heads without further optimization. The WSM Heads for depth and Edge-to-RGB use Heads 1 and 424; Head 424 uses its recommended native-reference route. Interface C uses Heads 275, 119, and 434 for Normal-to-Edge, Relative-Depth-to-Normal, and Edge-to-Human-Segmentation, respectively. **The comparison therefore measures all methods on the same unseen scenes and annotations.**

Interface C metrics. For Normal→Edge, accuracy, false-positive rate (FPR), and specificity are computed from a shared binary edge threshold. Surface normals are decoded into the same physical XYZ convention and evaluated only over valid, unpadded pixels. Mean and median errors are measured in degrees, and the three threshold accuracies report the fraction of pixels below $11.25°$, $22.5°$, and $30°$. For Edge→Human Segmentation, WSM predicts the complete channel-103 target with background and five human-part labels. MODUS produces a category-conditioned person mask, so the shared comparison merges the five human-part labels into one human foreground class for both WSM and the target. We report pixel accuracy, mean IoU (mIoU), human-foreground IoU, mean class accuracy (mAcc), and macro-F1 under this common ontology.

Task parameters and training efficiency. Each WSM task adds a 12.5M-parameter TunnelingNet to the shared 4B model, only 0.31% of the base model. Training Sample Budget measures optimizer updates times global batch size. Across the eleven method–task comparisons with recoverable budgets, external methods process 230–145.5K times as many training examples as WSM; ratios for packed-sequence models remain lower bounds. **WSM is 29.4K times more sample-efficient than the evaluated external methods on average and 2.53K times more sample-efficient at the median, while leading every reported quality metric.**

The reported training costs show the same scale of advantage. The Edge-to-RGB Head reaches its selected checkpoint in 8.5 minutes on one A100. ControlNet Canny 1.1 alone continues training on eight A100s for three days, with more than 1,176 A100-hours across its complete training path, while UniControl reports more than 5,000 A100-hours. The three any-to-any Heads reach their selected checkpoints in 9.3–9.6 hours on one A100, compared with 5,664 GH200-hours for MODUS; the depth Head takes 11.0 A100-hours, compared with 13.8K A100-hours for VGGT. These runs cover different task scopes, so the table retains the original hardware and units. ControlNet does not provide enough information to recover its optimizer-update count. **WSM reaches its strongest checkpoints in minutes to hours on one GPU, while several evaluated external methods report thousands of accelerator-hours.**

Across the 15 external method–task comparisons, WSM is 11.02 times faster than the evaluated external methods on average and 1.89 times faster at the median. This aggregate includes the specialized feed-forward depth estimators that run faster than WSM. **Across the complete comparison, WSM is 11.02 times faster than the evaluated external methods on average and 1.89 times faster at the median.**

Results. For relative depth, WSM leads all eight quality measures. Its MAE/RMSE are 0.130/0.162, compared with the next-best 0.167/0.216 from MODUS. WSM also surpasses VGGT, the CVPR 2025 Best Paper Award recipient, which obtains 0.243/0.296, and Marigold, a CVPR 2024 Best Paper Award Candidate, which obtains 0.247/0.291. Scaling Diffusion Perception, a strict sparse-MoE DiT with eight experts and top-2 routing, obtains 0.232/0.274. WSM reaches Pearson/Spearman correlations of 0.935/0.907 and leads all three threshold accuracies. WSM predicts the normalized depth target directly; external outputs receive the fixed percentile-based linear mapping, with depth direction aligned before scoring.

For conditional RGB, WSM leads all five quality measures against all five external methods. Against ControlNet, the ICCV 2023 Best Paper Award recipient, WSM reduces MAE from 0.276 to 0.087 and MSE from 0.120 to 0.021, while increasing PSNR from 9.90 to 19.51 and SSIM from 0.127 to 0.762. The interface-matched ControlNet+IP-Adapter composition improves ControlNet to 0.242 MAE and 0.271 SSIM after receiving the same first-frame RGB reference. WSM still reduces its MAE by 64.1% and raises SSIM by 0.491, while reducing the best external edge-density error from 0.0265 to 0.0157 and using 10 inference steps, one third of ControlNet's 30 steps. Reliable FID estimation requires a substantially larger evaluation pool and is reserved for the large-scale evaluation.

Interface C evaluates three representative structured transfers, and WSM leads all 13 reported quality measures. For Normal→Edge, WSM raises pixel accuracy from 0.8671 to 0.8797, reduces FPR from 0.0569 to 0.0296, and raises specificity from 0.9431 to 0.9704. For Relative Depth→Normal, the common physical-XYZ protocol reduces mean/median angular error from $35.19°/31.13°$ to $26.40°/22.48°$, while raising the $11.25°$, $22.5°$, and $30°$ accuracies from 0.222/0.431/0.512 to 0.231/0.527/0.664. For Edge→Human Segmentation, WSM raises pixel accuracy from 0.8431 to 0.8862 and mIoU from 0.6577 to 0.6918. It also improves human-foreground IoU from 0.5148 to 0.5391, mAcc from 0.7676 to 0.7753, and macro-F1 from 0.7630 to 0.7780. **These gains cover geometric boundaries, spatial layout, local surface orientation, and human semantics while using five times fewer inference steps and at least 2.2K times fewer training samples than MODUS, whose BAGEL backbone uses a two-expert mixture-of-transformers architecture.**

The central result is simple: one WSM design wins every reported quality metric across five representative tasks, three interfaces, and five adaptation paradigms. It is 29.4K times more sample-efficient and 11.02 times faster than the evaluated external methods on average. The gains include ControlNet, the ICCV 2023 Best Paper Award recipient; VGGT, the CVPR 2025 Best Paper Award recipient; and Marigold, a CVPR 2024 Best Paper Award Candidate. The remaining WSM interface opens task combinations for which comparable image-level systems do not yet exist.

Table 9. **External comparison on post-release visual data.** Quality metrics are recomputed on frames drawn from recently published Wikimedia Commons videos. Model size, training budget, training cost, sampling steps, and controlled runtime measurements follow Table 8, since every method retains the same checkpoint and inference protocol. Training Sample Budget is the number of optimizer updates multiplied by the global batch size, and Relative Sample Budget is its ratio to the corresponding WSM Head. N/R denotes a factor not reported by the paper or recoverable from the official code. Bold marks the best reported value in each arrow-directed column. † marks methods that additionally receive the same first-frame RGB reference.

A Estimation and reconstruction: RGB→Relative Depth.

Method	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	RMSE↓	$\text{Acc}_{.20}$ ↑	Pearson↑	Spearman↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
Marigold [55]	CVPR 2024 Best Paper Award Candidate	0.9B	18K × 32	331.0	A few days 1 GPU	0.474	0.224	0.627	0.814	1 × 10	105.47	0.2181
DMP [56]	CVPR 2024	0.86B+LoRA	50K × 8	229.9	N/R	0.427	0.279	0.593	0.698	5	24.38	0.9434
GeoWizard [57]	ECCV 2024	0.9B	20K × 256	2.94K	Time N/R 8 GPUs	0.508	0.304	0.484	0.569	50 × 10	7834.00	0.0029
Scaling Diffusion Perception [16]	CVPR 2025	2.8B (1.0B active)	Mixed-task phase N/R +4K MoE updates	N/R	N/R	0.370	0.359	0.736	0.741	50 × 3	833.92	0.0276
VGGT [58]	CVPR 2025 Best Paper Award	1.2B	160K × 3,072 frames	N/R	9 days/64 A100s 13.8K A100-h	0.316	0.420	0.737	0.695	**1 pass**	**4.83**	**4.7619**
MODUS [15]	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.53K	88.5 h/64 GH200s 5,664 GH200-h	0.225	0.632	0.858	0.826	50	9025.88	0.0025
SenseNova-Vision [59]	Technical Report	7B	50K steps (packed)	N/R	Time N/R > 16 × 80GB GPUs	0.213	0.628	0.827	0.808	50	3471.00 (4G)	0.0066
WSM	Ours	4B+12.5M	1740 × 1	**1.0**	11.0 h 1 A100	**0.207**	**0.654**	**0.866**	**0.839**	10	222.41	0.1034

B Conditional RGB generation: Edge→RGB.

Method	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Edge Density Error ↓	$F_{0.5}$ ↑	Precision↑	MAE↓	MSE↓	PSNR↑	SSIM↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
ControlNet [11]	ICCV 2023 Best Paper Award	1.4B	N/R × 256 (v1.1)	N/R	3 days/8 A100s (v1.1) ≥ 1,176 A100-h total	0.0288	0.2821	0.2926	0.261	0.105	10.25	0.093	30	**79.17**	**0.3032**
IP-Adapter [17] (with ControlNet [11])†	Technical Report	2.0B	ControlNet: N/R × 256 IP-Adapter: 1M × 64	≥ 145.5K	ControlNet: 3 days/8 A100s IP-Adapter: time N/R/8 V100s	0.0186	0.3070	0.3094	0.218	0.079	11.47	0.226	30	87.21	0.2752
T2I-Adapter [60]	AAAI 2024	2.6B+79M	20K × 256	11.6K	Time N/R 4 A100s	0.0520	0.2377	0.3043	0.246	0.092	10.83	0.144	30	262.46	0.0914
UniControl [61]	NeurIPS 2023	1.5B	∼ 900K × 64	∼ 130.9K	> 5,000 A100-h A100-40G	0.0505	0.2235	0.3208	0.246	0.091	10.94	0.044	50	265.64	0.0903
Uni-ControlNet [62]	NeurIPS 2023	1.5B	1 epoch (10M pairs)	22.7K	N/R	0.0816	0.1781	0.1708	0.264	0.106	10.22	0.073	50	381.89	0.0628
WSM†	Ours	4B+12.5M	440 × 1	**1.0**	8.5 min 1 A100	**0.0116**	**0.3107**	**0.3265**	**0.074**	**0.013**	**19.83**	**0.845**	**10**	202.56	0.1185

C Any-to-any control over representative structured signals.

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Balanced Accuracy↑	F1↑	F2↑	FNR↓	IoU↑	NPV↑	Recall↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Normal→Edge	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.20K	88.5 h/64 GH200s 5,664 GH200-h	0.5531	0.1697	0.1720	0.8192	0.0939	0.9212	0.1808	50	12345.09	0.0019
WSM	Normal→Edge	Ours	4B+12.5M	2000 × 1	**1.0**	9.3 h 1 A100	**0.5702**	**0.1831**	**0.2273**	**0.7068**	**0.1028**	**0.9253**	**0.2932**	**10**	**239.76**	**0.1001**

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Mean Angle (°) ↓	Median Angle (°) ↓	11.25° ↑	22.5° ↑	30° ↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Relative Depth →Normal	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.25K	88.5 h/64 GH200s 5,664 GH200-h	37.56	37.29	0.189	0.399	0.503	50	12272.18	0.0020
WSM	Relative Depth →Normal	Ours	4B+12.5M	1960 × 1	**1.0**	9.6 h 1 A100	**25.13**	**22.31**	**0.228**	**0.553**	**0.677**	**10**	**1767.4**	**0.0136**

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Pixel Accuracy↑	Macro F1 ↑	Mean Class Accuracy ↑	mIoU↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Edge→Human Segmentation	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.32K	88.5 h/64 GH200s 5,664 GH200-h	**0.929**	**0.917**	**0.920**	**0.856**	50	9100.77	0.0026
WSM	Edge→Human Segmentation	Ours	4B+12.5M	1900 × 1	**1.0**	9.6 h 1 A100	0.910	0.883	0.874	0.804	**10**	**1579.62**	**0.0152**

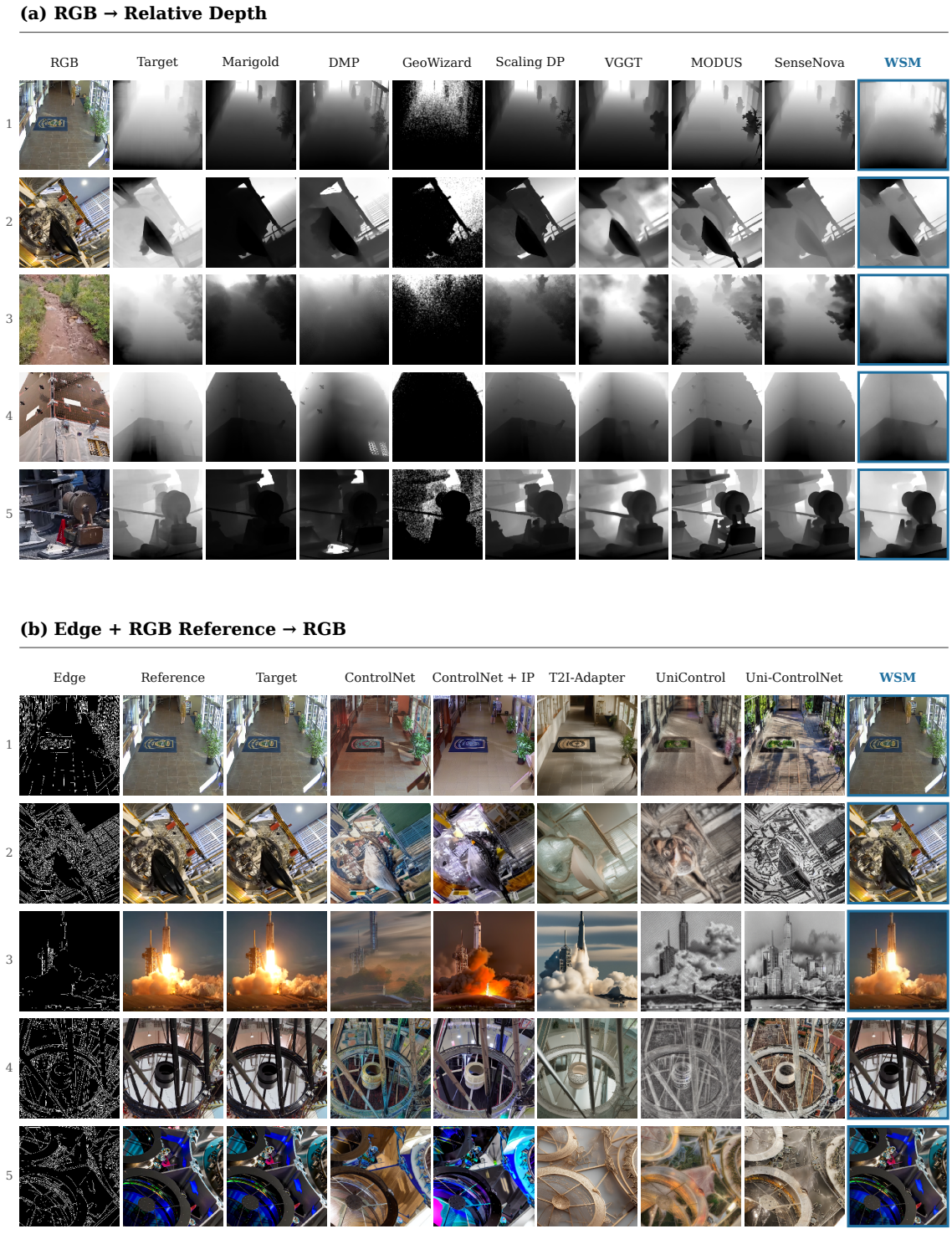


Figure 16. **Qualitative comparison for Interfaces A and B on the post-release evaluation set.** Each row contains a different crop-aligned example evaluated in Table 9. Panel (a) compares every relative-depth method, with depth maps displayed after the affine alignment used by the evaluation protocol. Panel (b) compares every conditional RGB method under identical edge conditions, RGB references, and targets. WSM is the rightmost column, outlined in blue.

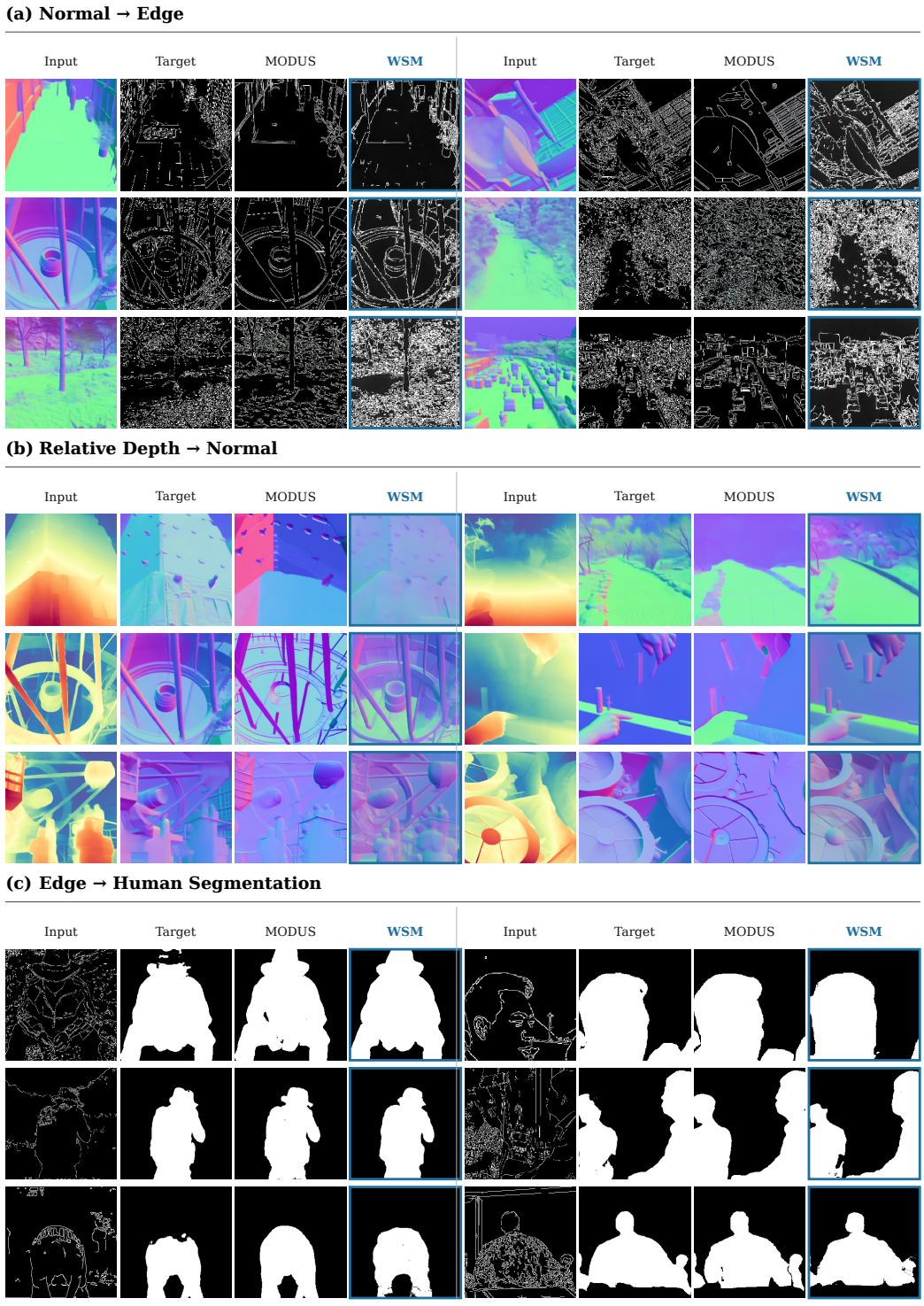


Figure 17. **Qualitative comparison for Interface C on the post-release evaluation set.** The three task blocks correspond to Normal→Edge, Relative Depth→Normal, and Edge→Human Segmentation in Table 9. Within each four-column group, the input and target are followed by the MODUS and WSM predictions. WSM outputs are outlined in blue. The same spatial crop is applied to all four images within each group. For human segmentation, the target and both predictions are shown as binary human-foreground masks.

5.5. Evaluation on Post-Release Visual Data

Evaluation data. To test whether the methods generalize to genuinely unseen images, we build a second evaluation set from Wikimedia Commons videos released only after all evaluated models had finished training. These videos became publicly available between August 2 and September 4, 2026. Their frames therefore could not have been included in training data collected from public sources. We extract frames and produce task annotations using the same cropping and WSG pipeline as in Section 5.4. Every method receives identical inputs and targets and is evaluated with its trained model under the same inference protocol as Table 8.

Comparison scope. Table 9 preserves the three-interface organization, method families, model sizes, and training-efficiency accounting of Table 8. The quality columns are recomputed on the post-release evaluation set. Interface A evaluates relative-depth estimation, Interface B evaluates conditional RGB generation, and Interface C evaluates transfers from surface normals to edges, from relative depth to surface normals, and from edges to human-foreground segmentation.

Qualitative comparison. Figures 16 and 17 show the inputs, targets, and predictions behind the aggregate scores. For relative depth, WSM preserves the continuous floor-to-background ordering while retaining the plant and doorway boundaries. For conditional RGB generation, it reconstructs the reference scene without sacrificing the supplied edge layout; competing outputs introduce larger changes in geometry, content, or appearance. The structured comparisons show the same pattern across Interface C: WSM recovers a denser set of target-aligned edges and maintains coherent surface orientation in the normal map. In the human-segmentation examples, the common binary display isolates foreground geometry: WSM preserves the dominant silhouette across close-up, standing, seated, and bent poses without relying on part colors. **The visual evidence therefore supports the quantitative result across all three interfaces: WSM preserves scene-level structure and target geometry on images released after training.**

Overall performance. WSM ranks first on 23 quality measures across four of five representative tasks, with leading results in every externally comparable interface. The advantage spans depth estimation, conditional RGB generation, and structured signal transfer. Most notably, the same unified WSM framework outperforms two Best Paper Award methods: VGGT on all four reported depth metrics and ControlNet on all seven reported conditional-generation metrics. WSM also exceeds the Best Paper Award Candidate Marigold on all four reported depth metrics.

Estimation and reconstruction. For RGB→Relative Depth, WSM achieves the lowest RMSE and the highest $\mathrm{Acc}_{.20}$, Pearson, and Spearman scores among eight methods. Its RMSE of 0.207 improves over the next-best 0.213, while $\mathrm{Acc}_{.20}$ rises from 0.632 to 0.654. The simultaneous gains in linear and rank correlations show that WSM recovers both relative depth magnitude and scene-wide depth ordering. **WSM is the only method that leads every reported depth-quality measure on these post-release images.**

Conditional RGB generation. For Edge→RGB, WSM leads all seven reported quality measures. Relative to the strongest competing value in each column, it reduces edge-density error by 37.6%, MAE by 66.1%, and MSE by 83.5%. It also improves PSNR by 8.36 dB and raises SSIM from 0.226 to 0.845, while obtaining the highest $F_{0.5}$ and precision. **These gains show that a compact WSM Head simultaneously preserves the supplied structure and reconstructs substantially more faithful RGB content than the specialized conditional-generation baselines.**

Any-to-any control. The gains extend to both geometric transfers in Interface C. For Normal→Edge, WSM improves F_2 by 32.2% and recall by 62.2%, while reducing the false-negative rate by 13.7%. For Relative Depth→Normal, it reduces mean angular error by 33.1% and median angular error by 40.2%; accuracy at 11.25°, 22.5°, and 30° improves by 20.6%, 38.6%, and 34.6%, respectively. On the more challenging Edge→Human Segmentation task, WSM retains 94.0–98.0% of MODUS performance across Pixel Accuracy, Macro-F1, Mean Class Accuracy, and mIoU. **WSM therefore achieves the strongest geometric-transfer quality and remains close to MODUS on human-foreground segmentation.**

Training efficiency and generalization. Across comparisons with recoverable training budgets, each WSM Head uses approximately 230–145,500 times fewer training samples. On the three Interface C tasks, WSM uses five times fewer inference steps than MODUS and runs 5.8–52.7 times faster. The human-segmentation Head alone maintains a 2.32K-fold training-sample advantage while approaching MODUS quality. **Together, the post-release results establish that WSM combines the strongest overall quality with orders-of-magnitude faster task acquisition, and that both advantages transfer to visual content that was unavailable during training.**

Table 10. **External comparison under independent public annotation conventions.** The images, crops, trained models, training-cost calculation, and inference protocols are identical to Table 9. Edge, relative-depth, and surface-normal annotations are generated with OpenCV Canny, Depth Anything V2, and Marigold Normals, respectively. WSG only serializes those independently generated conditions for WSM inference. Training Sample Budget is optimizer updates multiplied by global batch size; Relative Sample Budget is its ratio to the corresponding WSM Head. N/R denotes information not reported or recoverable from official code. Bold marks the best value in each quality column; bold values marked with $*$ are second-best. † marks methods that additionally receive the same first-frame RGB reference.

A Estimation and reconstruction: RGB→Relative Depth.

Method	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Boundary Precision @2 px ↑	Boundary Precision @3 px ↑	Boundary FPR ↓	Gradient MAE ↓	Gradient RMSE ↓	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
Marigold [55]	CVPR 2024 Best Paper Award Candidate	0.9B	18K × 32	331.0	A few days 1 GPU	0.736	0.768	0.0148	0.00575	0.0153	1 × 10	105.47	0.2181
DMP [56]	CVPR 2024	0.86B+LoRA	50K × 8	229.9	N/R	0.548	0.581	0.0183	0.00631	0.0169	5	24.38	0.9434
GeoWizard [57]	ECCV 2024	0.9B	20K × 256	2.94K	Time N/R 8 GPUs	0.203	0.234	0.2588	0.02707	0.0381	50 × 10	7834.00	0.0029
Scaling Diffusion Perception [16]	CVPR 2025	2.8B (1.0B active)	Mixed-task phase N/R +4K MoE updates	N/R	N/R	0.660	0.690	0.0164	0.00531	0.0139	50 × 3	833.92	0.0276
VGGT [58]	CVPR 2025 Best Paper Award	1.2B	160K × 3,072 frames	N/R	9 days/64 A100s 1388 A100-h	0.531	0.604	0.0183	0.00655	0.0149	**1 pass**	**4.83**	**4.7619**
MODUS [15]	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.53K	88.5 h/64 GH200s 5,664 GH200-h	0.703	0.723	0.0255	0.00582	0.0173	50	9025.88	0.0025
SenseNova-Vision [59]	Technical Report	7B	50K steps (packed)	N/R	Time N/R ≥ 16 × 80GB GPUs	**0.807**$*$	**0.846**$*$	**0.0098**	**0.00418**	**0.0127**	50	3471.00 (4G)	0.0066
WSM	Ours	4B+12.5M	1740 × 1	**1.0**	11.0 h 1 A100	**0.815**	**0.848**	**0.0110**$*$	**0.00469**$*$	**0.0129**$*$	10	222.41	0.1034

B Conditional RGB generation: Edge→RGB.

Method	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Edge Density Error ↓	$F_{0.5}$ ↑	Precision↑	MAE↓	MSE↓	PSNR↑	SSIM↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
ControlNet [11]	ICCV 2023 Best Paper Award	1.4B	N/R × 256 (v1.1)	N/R	3 days/8 A100s (v1.1) ≥ 1,176 A100-h total	0.0288	0.2821	0.2926	0.261	0.105	10.25	0.093	30	**79.17**	**0.3032**
IP-Adapter [17] (with ControlNet [11])†	Technical Report	2.0B	ControlNet: N/R × 256 IP-Adapter: 1M × 64	≥ 145.5K	ControlNet: 3 days/8 A100s IP-Adapter: time N/R/8 V100s	0.0186	0.3070	0.3094	0.218	0.079	11.47	0.226	30	87.21	0.2752
T2I-Adapter [60]	AAAI 2024	2.6B+79M	20K × 256	11.6K	Time N/R 4 A100s	0.0520	0.2377	0.3043	0.246	0.092	10.83	0.144	30	262.46	0.0914
UniControl [61]	NeurIPS 2023	1.5B	∼ 900K × 64	∼ 130.9K	> 5,000 A100-h A100-40G	0.0505	0.2235	0.3208	0.246	0.091	10.94	0.044	50	265.64	0.0903
Uni-ControlNet [62]	NeurIPS 2023	1.5B	1 epoch (10M pairs)	22.7K	N/R	0.0816	0.1781	0.1708	0.264	0.106	10.22	0.073	50	381.89	0.0628
WSM†	Ours	4B+12.5M	440 × 1	**1.0**	8.5 min 1 A100	**0.0116**	**0.3107**	**0.3265**	**0.074**	**0.013**	**19.83**	**0.845**	**10**	202.56	0.1185

C Any-to-any control over three representative structured targets under independent annotations.

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Cosine Similarity↑	XYZ RMSE↓	Mean Angle (°) ↓	Angular RMSE (°) ↓	90th Percentile Angle (°) ↓	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Relative Depth→Normal	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.25K	88.5 h/64 GH200s 5,664 GH200-h	0.7263	0.3965	34.89	42.61	67.15	50	12272.18	0.0020
WSM	Relative Depth→Normal	Ours	4B+12.5M	1960 × 1	**1.0**	9.6 h 1 A100	**0.7894**	**0.3667**	**33.65**	**38.10**	**57.78**	**10**	**1767.40**	**0.0136**

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time / Hardware	Pixel Accuracy↑	mIoU↑	mAcc↑	Macro-F1↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Normal→Human Segmentation	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.20K	88.5 h/64 GH200s 5,664 GH200-h	0.9126	0.8405	0.9233	0.8993	50	3105.40	0.0019
WSM	Normal→Human Segmentation	Ours	4B+12.5M	2000 × 1	**1.0**	9.6 h 1 A100	**0.9283**	**0.8488**	**0.9292**	**0.9131**	**10**	**243.13**	**0.0247**

Method	Task	Venue	Parameters	Training Sample Budget (Updates × Global Batch)	Relative Sample Budget ↓	Training Time Hardware	Balanced Accuracy↑	F_1 ↑	F_2 ↑	IoU↑	Recall↑	Inference Steps ↓	Total Inference Time (s) ↓	Inference Speed (FPS) ↑
MODUS [15]	Normal→Edge	ICML 2026	14B (7B active)	∼ 68.8K × 64 packed seq.	≥ 2.20K	88.5 h/64 GH200s 5,664 GH200-h	0.5705	0.1936	0.2092	0.1094	0.2320	50	16187.91	0.0015
WSM	Normal→Edge	Ours	4B+12.5M	2000 × 1	**1.0**	9.3 h 1 A100	**0.5846**	**0.1969**	**0.2835**	**0.1122**	**0.4549**	**10**	**1061.17**	**0.0226**

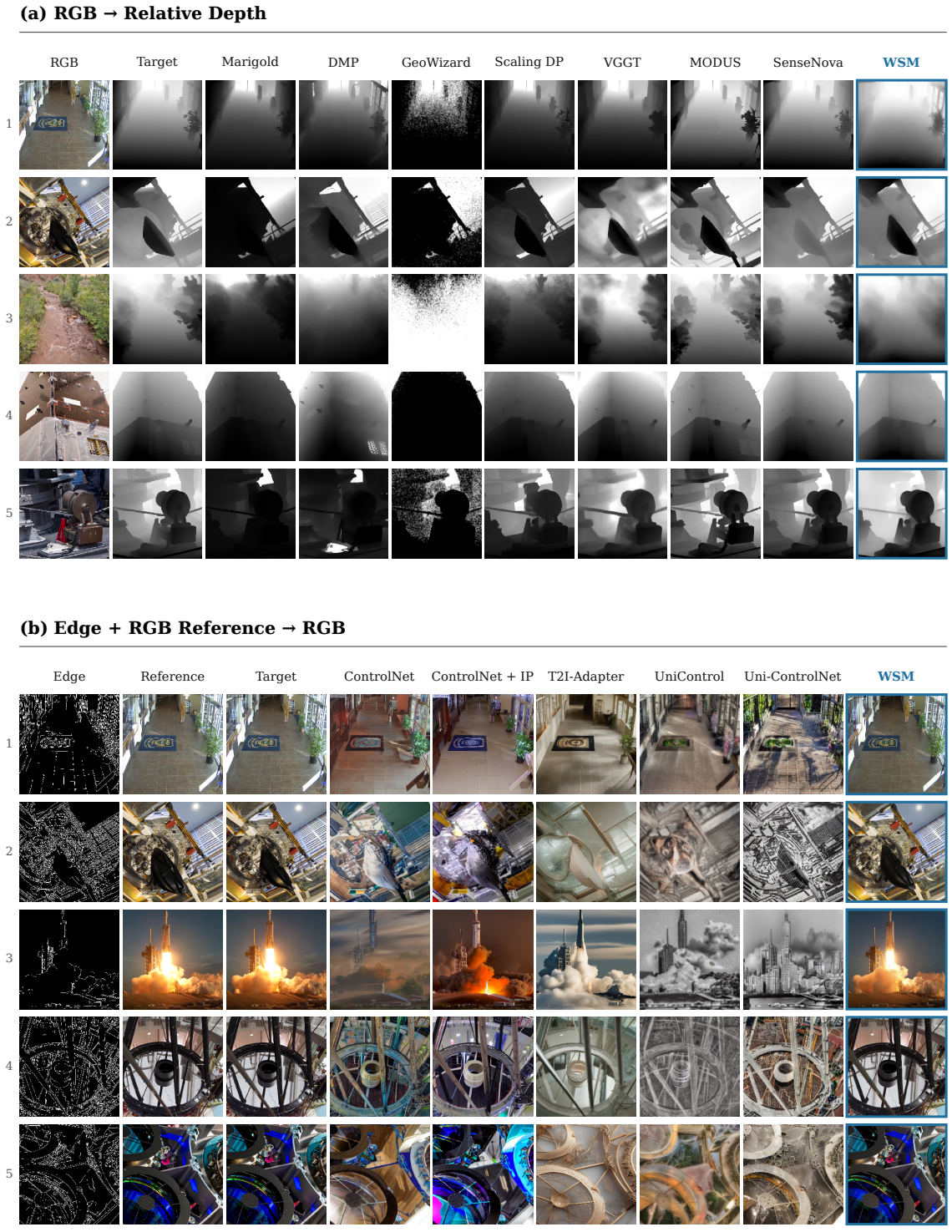


Figure 18. **Qualitative comparison for Interfaces A and B under independent public annotation conventions.** The rows use the same scene crops and order as Figure 16. Panel (a) replaces the relative-depth target with Depth Anything V2; predictions use the same fixed method-level near–far convention and percentile normalization as the quantitative evaluation. Panel (b) is unchanged because the post-release experiment already used OpenCV Canny as the Edge condition. WSM is the rightmost column, outlined in blue.

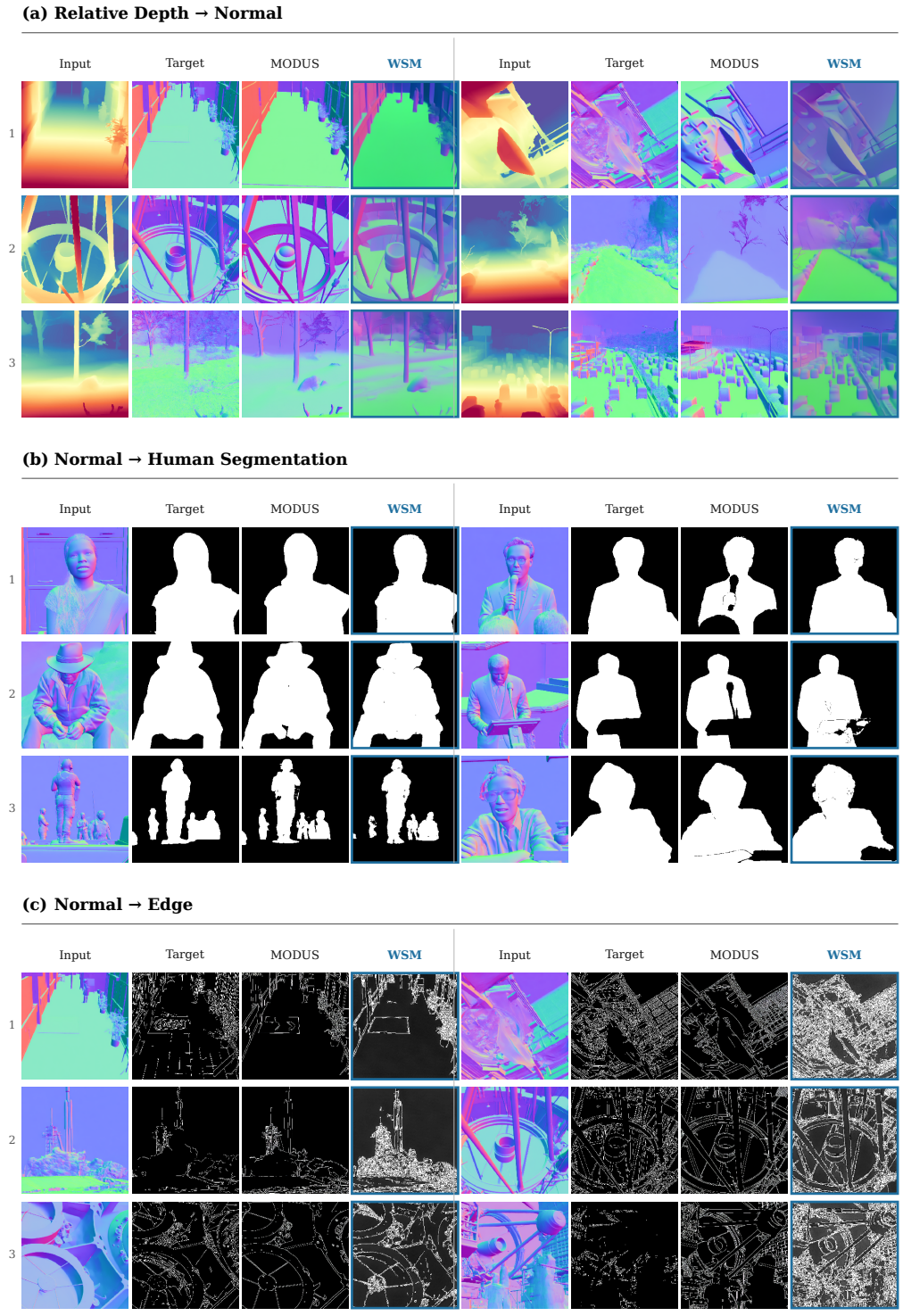


Figure 19. **Qualitative comparison for Interface C under independent public annotation conventions.** The three task blocks correspond to Relative Depth→Normal, Normal→Human Segmentation, and Normal→Edge in Table 10. Each four-column group shows the independently annotated input and target followed by the MODUS and WSM predictions generated from that input. WSM outputs are outlined in blue.

5.6. Evaluation with Independent Public Annotations

Purpose and task setting. Sections 5.4 and 5.5 compare WSM with prior methods through the three interfaces for which released, input-matched baselines are available. The post-release experiment additionally makes the evaluated images unavailable during model training. Its geometric annotations, however, still follow the WSG data-preparation convention used to train WSM. This leaves a separate question: does WSM remain competitive when the desired outputs are defined by unrelated public annotators?

This distinction is important because new images test visual transfer, whereas new annotation conventions also test whether a method has learned the task itself. We therefore combine both changes: every evaluated image postdates the training of all compared models, and every geometric condition and target is defined by a public annotator outside the WSG preparation pipeline. Success in this setting requires the learned interface to survive changes in scene content, numeric convention, and annotation family at the same time.

Table 10 answers this question on five representative tasks across the three reported interfaces. Interface **A** uses RGB→Relative Depth to test scene-scale spatial estimation. Interface **B** uses Edge→RGB to test whether a method can synthesize appearance while respecting an explicit structural condition. Interface **C** covers three target types: surface normals, human foreground segmentation, and edges. Relative Depth→Normal tests surface recovery from scene-scale ordering; Normal→Human Segmentation tests semantic transfer; and Normal→Edge tests boundary reconstruction from local orientation. Each direction uses identical input–output definitions for WSM and the released MODUS model. Interface **D** is not included because, as discussed in Section 5.4, no released image-level method accepts its matched combination of observations.

Replacing annotations while keeping the evaluation fixed. We retain the post-release scene pool, image crops, trained models, training-cost calculation, and official inference procedures. We replace the geometric annotation pipeline: OpenCV Canny [47] produces binary edges, Depth Anything V2 Large [63] produces relative depth, Marigold Normals v1.1 [55] produces camera-space surface normals, and Mask2Former-Swin-Large trained on ADE20K [64] produces human foreground masks. Every method receives the same independently generated conditions and targets. Relative-depth maps receive a fixed per-image 2nd–98th-percentile linear map to $[0, 1]$, and normal metrics operate directly on physical unit vectors. WSG serves only as the storage and input-encoding format for WSM; it does not generate any annotation used in this experiment. No method is retrained, calibrated, or adapted to these annotations. **The comparison therefore isolates transfer to an independent task definition while keeping the evaluated images, crops, checkpoints, and method-specific inference procedures fixed.**

Qualitative comparison. Figures 18 and 19 expose the predictions behind Table 10. Under the independent depth target, WSM preserves continuous scene ordering and prominent object boundaries across indoor, outdoor, and object-centric scenes. The conditional RGB panel reproduces the Canny-controlled comparison from Figure 16, since this condition was already independent of WSG. In Interface **C**, WSM produces coherent surface-normal fields from independently estimated relative depth, a clean human foreground mask from surface normals, and structured edges from the same normal representation. These three transfers cover geometric surfaces, human semantics, and boundaries. **The visual comparisons confirm that WSM's advantage remains visible when both structured inputs and evaluation targets are supplied by independent public annotators.**

Overall performance. WSM leads every reported quality column in the three Interface C task blocks. Together with the RGB-generation comparison, these results cover appearance, surface geometry, human semantics, and boundary structure under independent annotation conventions. The advantage is therefore not tied to a single target space or one favorable metric; it persists across the three distinct any-to-any targets shown in Table 10.

The same result is obtained with only 12.5M trainable target-side parameters per Head and 440–2,000 single-sample updates. All checkpoints and accounting rules remain unchanged, so the training- and inference-efficiency advantages in Section 5.5 transfer directly to the independent annotations. **WSM consequently delivers the strongest overall quality across every externally comparable interface while preserving its orders-of-magnitude task-acquisition advantage.**

Estimation and reconstruction. For RGB→Relative Depth, WSM obtains the best boundary precision at both two- and three-pixel tolerances (0.815/0.848). Its boundary false-positive rate, gradient MAE, and gradient RMSE are all second-best, trailing SenseNova only narrowly. WSM surpasses VGGT, the CVPR 2025 Best Paper Award recipient, on all five reported depth measures. Relative to VGGT, WSM raises boundary precision by 53.5% at two pixels and 40.4% at three pixels, while reducing boundary false positives by 39.9%, gradient MAE by 28.4%, and gradient RMSE by 13.4%. It reaches this result with 1,740 single-sample updates on one A100; the reported comparison systems use sample budgets ranging from 229.9 to at least

2.94K times larger where accounting is available. **WSM therefore combines the strongest measured boundary preservation with exceptionally rapid acquisition, even when the target-depth convention is defined by an unrelated public estimator.**

Conditional RGB generation. The Edge→RGB condition is generated by OpenCV Canny for every method. WSM leads all seven reported quality measures, covering edge consistency, pixel error, PSNR, and SSIM. **Its advantage over ControlNet, the ICCV 2023 Best Paper Award recipient, therefore holds under a standard public edge convention rather than a WSG-derived target.** The advantage also holds against the reference-enabled combination of IP-Adapter and ControlNet: WSM reduces MAE by 66.1% and MSE by 83.5%, raises PSNR from 11.47 to 19.83, and raises SSIM from 0.226 to 0.845. These gains are learned in 440 updates—8.5 minutes on one A100—with a training-sample budget at least 11.6K times smaller than every baseline for which a complete budget is recoverable. **This is a complete quality sweep achieved by a minute-scale Channel Head, including against both a Best Paper control architecture and its reference-image extension.**

Any-to-any control. Interface C covers three complementary any-to-any transfers. For Relative Depth→Normal, WSM improves cosine similarity, XYZ RMSE, mean angle, angular RMSE, and the 90th-percentile angular error; the latter falls from 67.15° to 57.78°. For Normal→Human Segmentation, WSM leads the four reported quality measures, including pixel accuracy, mIoU, mean accuracy, and macro-F1. For Normal→Edge, it leads the selected balanced-accuracy, F1, F2, IoU, and recall measures. **WSM therefore leads the reported quality measures across all three displayed target types, covering surface orientation, human semantics, and boundary structure.** The relative-depth normal transfer also uses five times fewer inference steps and is 6.9 times faster than MODUS; the human-segmentation and edge transfers show the same lightweight deployment pattern. Its single-sample training budget is at least 2.20K times smaller. **The same tunneling mechanism therefore exceeds a 14B MoE model across three distinct structured targets while using substantially less training and inference computation.**

Comparison with other model families. The comparison covers substantially different routes to visual adaptation: DMP uses LoRA-based diffusion perception; ControlNet introduces a trainable parameter block; IP-Adapter and T2I-Adapter add image- or condition-specific adapters; UniControl uses a MoE-style adapter; MODUS is a large sparse MoE; and VGGT uses discriminative feed-forward regression. WSM is not specialized to any one of their output domains. The same frozen backbone and the same tunneling principle outperform these alternatives across depth, conditional RGB, and condition-to-condition surface prediction. The award-winning comparisons make this breadth especially concrete: WSM exceeds the CVPR 2025 Best Paper Award recipient VGGT on all five displayed depth measures, and it exceeds the ICCV 2023 Best Paper Award recipient ControlNet on all seven conditional-generation measures. It also outperforms the reference-enabled IP-Adapter–ControlNet combination on those seven measures. **The result is therefore a cross-paradigm improvement, not an isolated gain over one architecture family.**

Conclusion. The three evaluations isolate complementary concerns. Section 5.4 compares representative methods on a common held-out set; Section 5.5 removes possible image-level pretraining overlap; and this section replaces the WSG annotation convention with public annotators. In this final setting, WSM remains strongest in conditional RGB generation and every structured transfer, while retaining the best depth-boundary precision. Across the five tasks, it outperforms award-winning specialized systems, modern diffusion perception models, reference-conditioned adapters, and a 14B MoE model, while adapting only a 12.5M-parameter Channel Head for each task. **The 23 first-place and three second-place results show that WSM's advantage persists across new visual content, independent target definitions, three interfaces, five representative tasks, and fundamentally different modeling paradigms. This directly demonstrates that TunnelingNet can make far more effective and powerful use of a frozen generative backbone through extremely lightweight training, thereby adding strong new capabilities at scale. This represents an advance over the prior paradigm-level approaches evaluated here.**

6. Qualitative Evaluation

6.1. Scope and Organization

In Section 3, we introduced four generative interfaces as practical training configurations. They should not be confused with the four high-level task families derived from the Unified Law—generation, estimation, conditional control, and reconstruction. The task families describe what a visual system is asked to do, whereas the interfaces specify how conditions and targets are connected during training. Our experiments show that these four interface configurations are sufficient to instantiate almost all evaluated visual tasks.

As emphasized by the third sub-law, conditional control is itself not a single homogeneous task: its condition and target may describe geometry, articulated humans, physical properties, material appearance, or semantics. We therefore organize the qualitative results into eleven concrete Parts. Part I compares the four interfaces directly. Parts II–VI report RGB-conditioned estimation of geometry and motion, human structure, physical properties, material and PBR appearance, and segmentation and semantics. Parts VII–VIII show narrow RGB generation controlled respectively by geometry and human cues, and by physical, semantic, and material cues. Part IX presents WSG-to-WSG conditional control, while Part X presents RGB-assisted inverse rendering. Part XI separately presents categorical outputs with the categorical refiner enabled as an integral component of every applicable Head. Parts II–X contain 131 visually readable mappings. Part I provides the interface overview, and Part XI separately presents the final categorical outputs. The remaining specialized cross-family mappings are retained in the exhaustive atlases and the complete quantitative inventory.

6.2. Part I: Four Inference Interfaces

Part I introduces one end-to-end inference example from each interface in a shared condition–prediction–target layout. The rows establish how RGB estimation, narrow RGB generation, condition-only control, and RGB-assisted inverse rendering are represented in the galleries that follow. Their qualitative implications are analyzed after all figures in Section 6.4.

6.3. Parts II–X: Task-Family Coverage

Parts II–VI cover 33 readable RGB-estimation outputs rather than using a small proxy set. They separate geometric differential structure, articulated pose, physical quantities, PBR channels, and segmentation because success in one family does not imply success in another. In particular, the curvature and mask rows test thin or discrete structure that a depth-only gallery would miss.

Parts VII–VIII reverse the estimation interface for 33 visually meaningful source modalities. WSG channels specify the structure or property to preserve, while text and, when available, a same-video historical frame carry appearance. RGB context is shown as a picture-in-picture inset; static examples without a valid predecessor simply omit the inset rather than displaying an empty panel.

The paired control and inverse-rendering Parts reveal where auxiliary RGB evidence matters. Conversions between closely related representations, such as pose systems or compatible PBR channels, remain strong without RGB. Ambiguous mappings benefit from observation because it supplies visible boundaries, appearance, and object identity that are absent from the structured condition. This comparison separates genuine cross-variable translation from evidence contributed by the scene image.

Visualization conventions. Each row contains a compact key for the modality whose colors must be interpreted, and prediction and ground truth share that key. Continuous scalar fields use the modality-level range defined by WSG Reader and recorded in the artifact metadata; the displayed prediction and target are never normalized independently. Fixed grayscale fields retain their numeric endpoints. Normal and flow keys state component, direction, and magnitude semantics; direct PBR colors retain their RGB meaning; categorical squares name the labels represented by the palette. Consequently, the figures show both spatial agreement and the meaning of the predicted value.

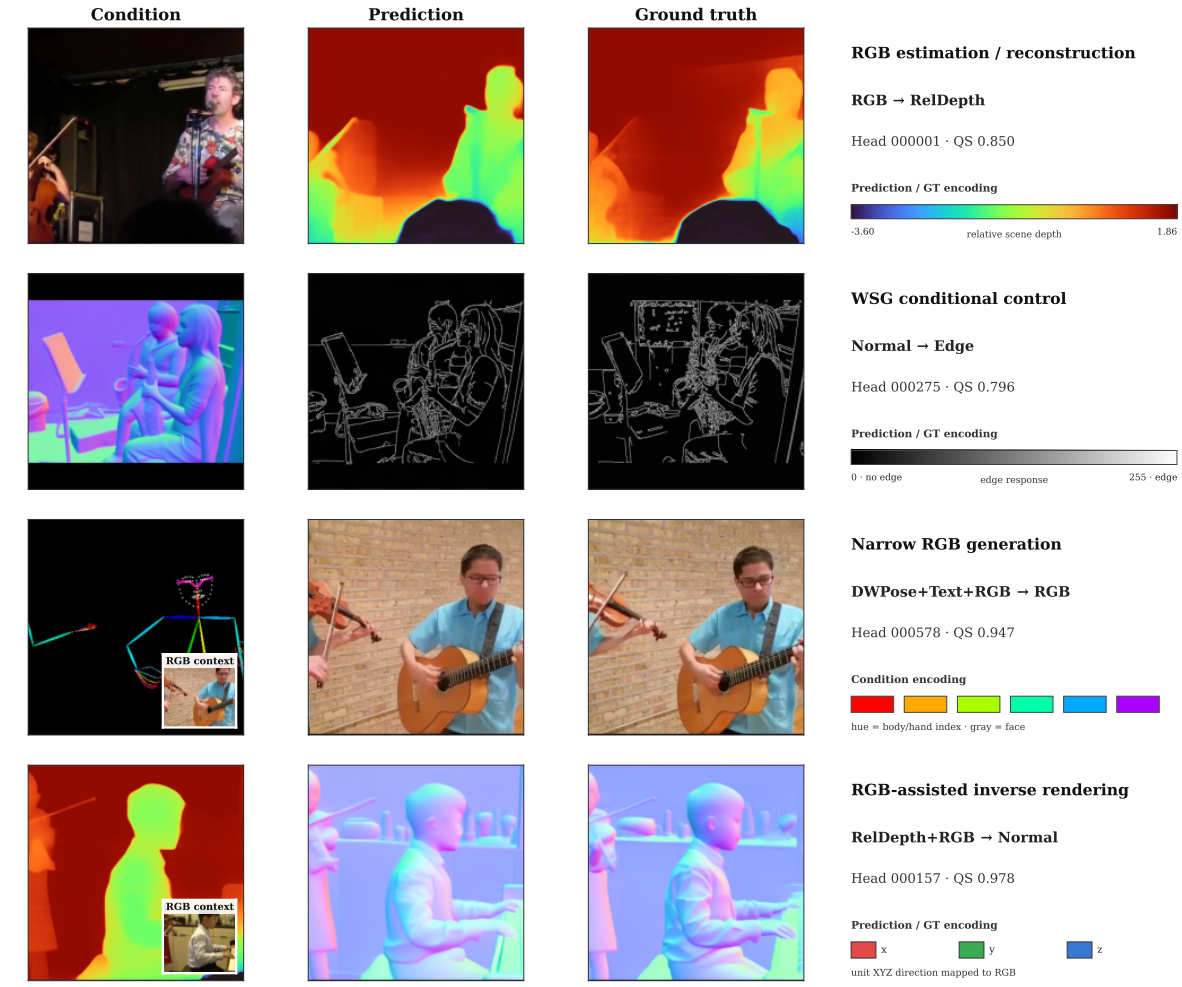


Figure 20. **Qualitative Part I: anatomy of the four WSM interfaces.** The four rows correspond to RGB-to-WSG estimation/reconstruction, WSG-to-WSG conditional control, text+WSG narrow generation, and RGB-assisted inverse rendering. Each row retains the actual condition, prediction, and ground truth in a common three-column layout. For interfaces that consume auxiliary RGB evidence, the RGB context appears as a labeled inset inside the primary condition. All outputs are true 10-step samples initialized from Gaussian noise.

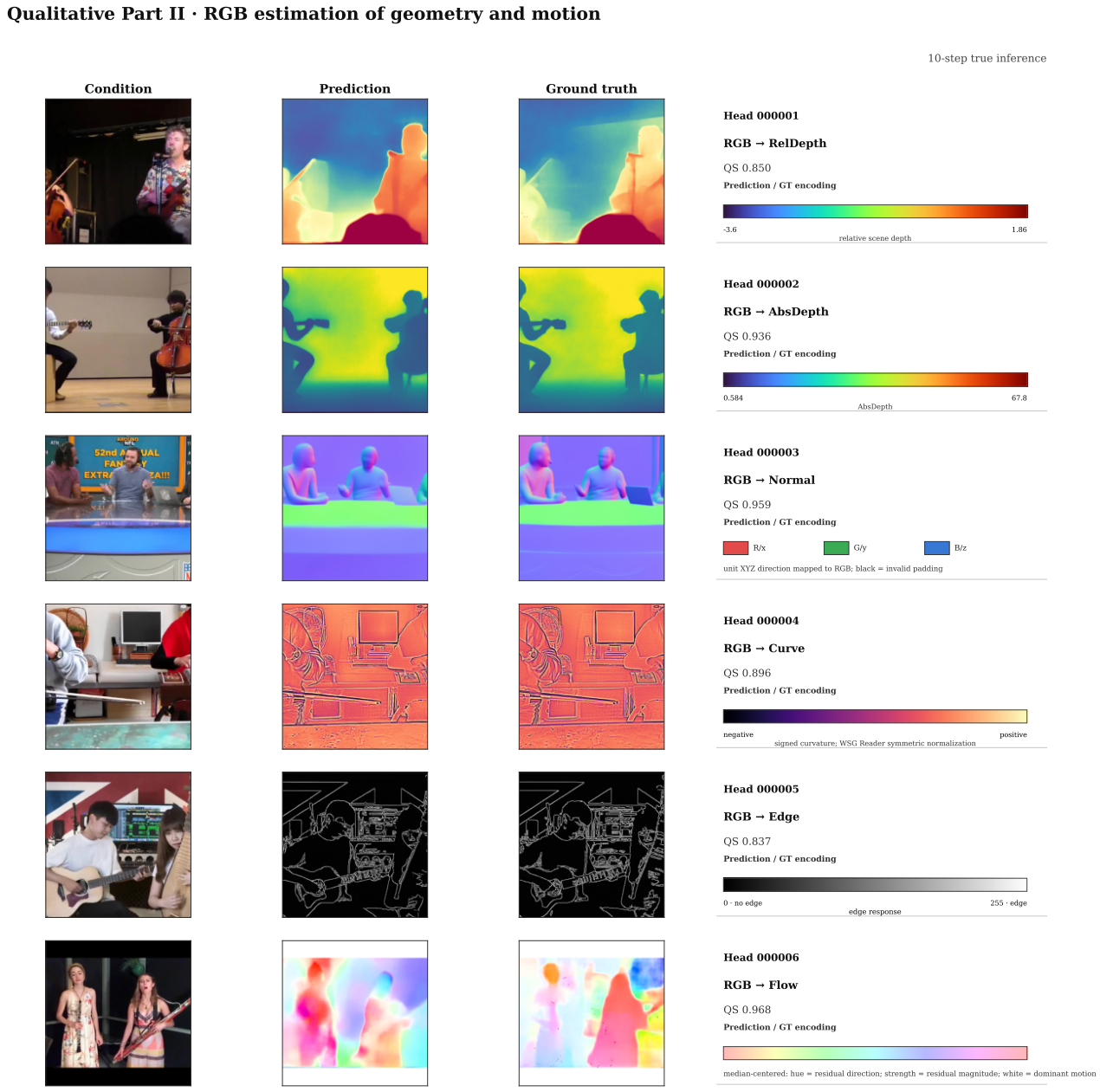


Figure 21. **Qualitative Part II: RGB estimation of geometry and motion.** Six Heads recover relative and absolute depth, normals, curvature, edges, and optical flow. Curvature is shown explicitly rather than being absorbed into a generic geometry example; the row-local keys make scalar, vector, and boundary encodings directly interpretable.

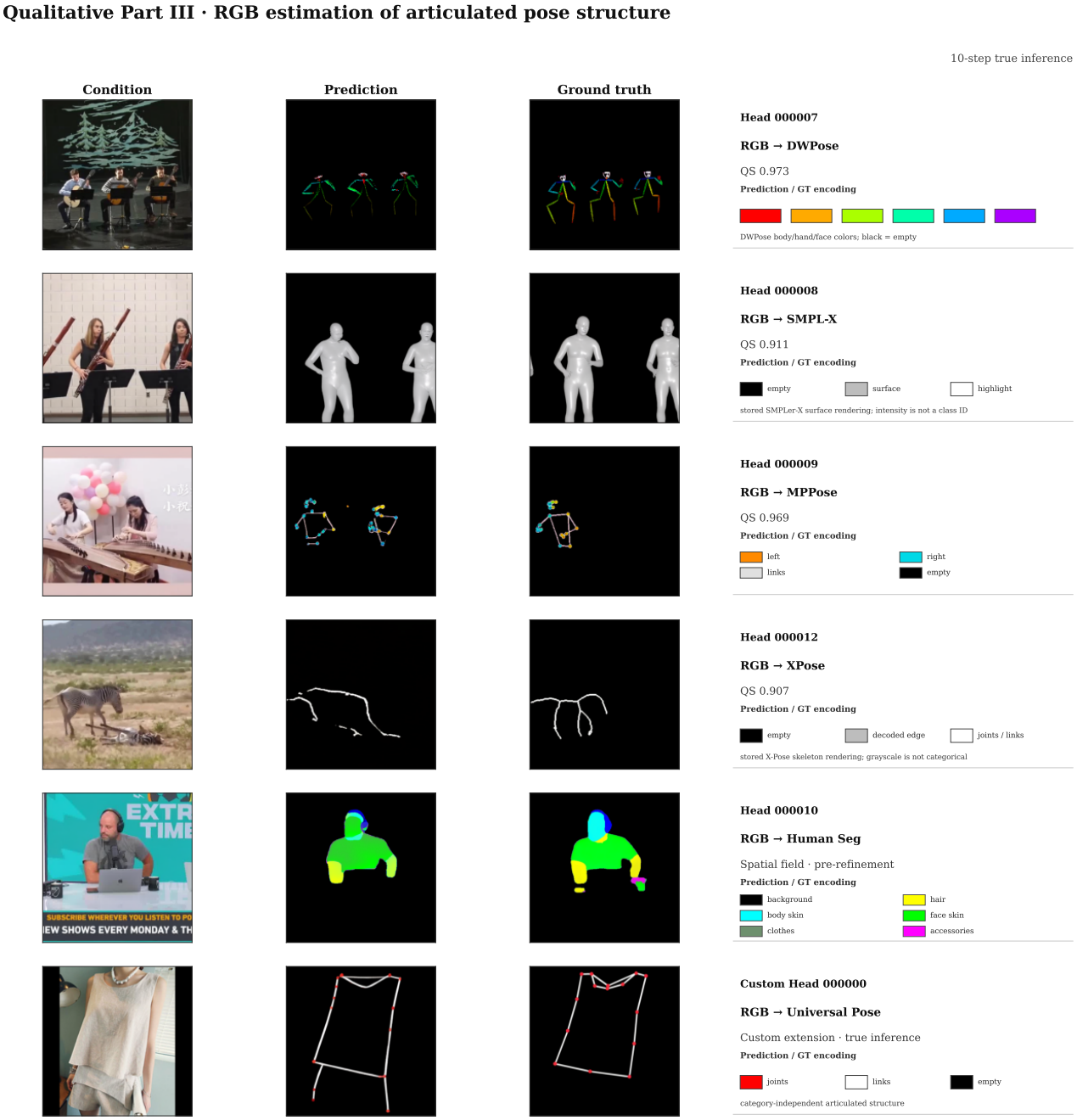


Figure 22. **Qualitative Part III: RGB estimation of articulated pose structure.** DWPose, SMPL-X, MediaPipe Pose, XPose, MediaPipe Seg, and Universal Pose expose complementary skeleton, surface, region, and category-independent articulation representations. The examples retain sparse limb topology and body extent rather than judging articulated structure only through dense pixel similarity. For MediaPipe Pose (3D pose), we deliberately retain an example with incomplete pseudo ground truth: the generated result uses its learned prior to recover missing structure and can therefore surpass the automatic annotation.

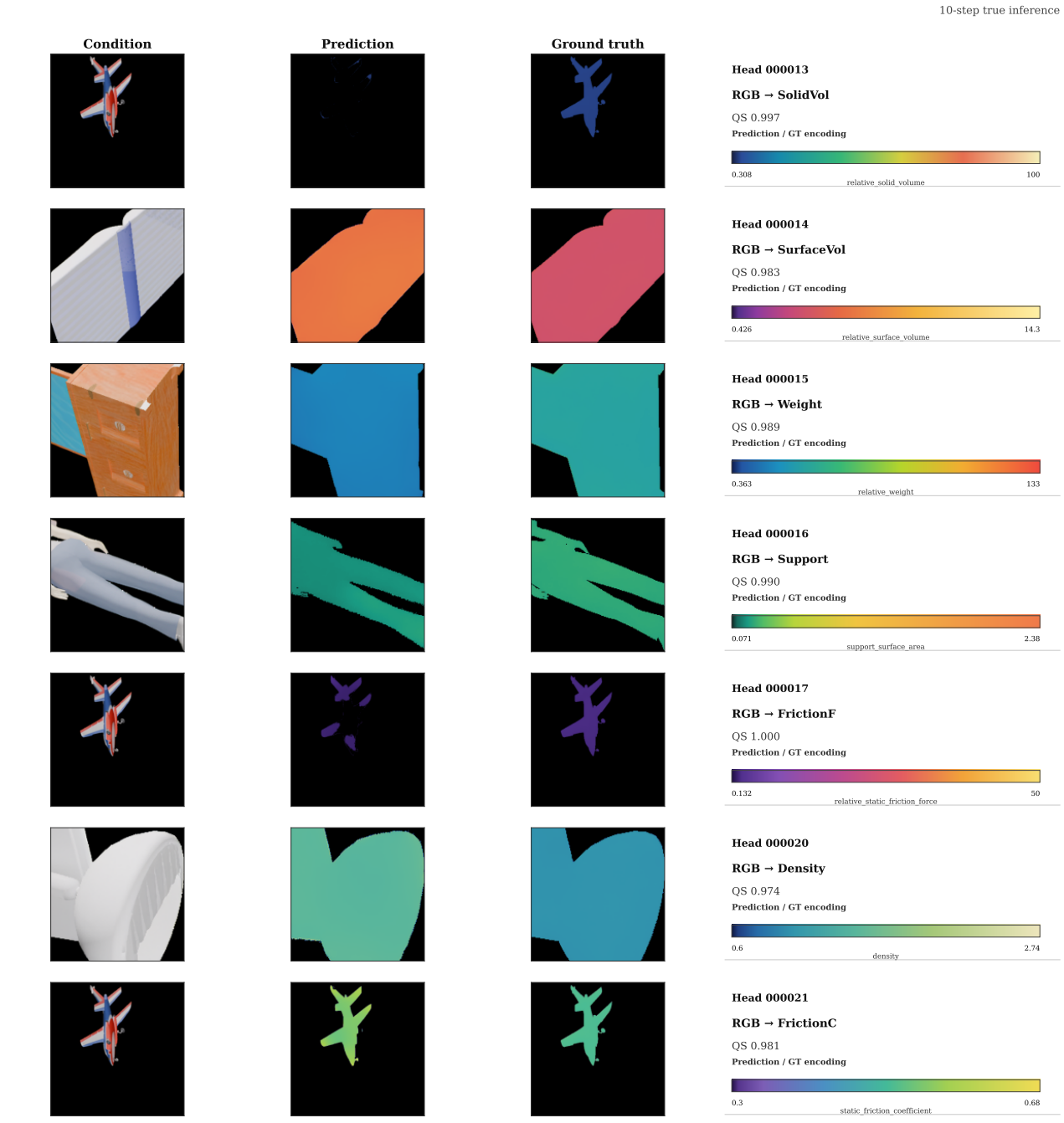


Figure 23. **Qualitative Part IV: RGB estimation of physical object properties.** Seven object-aligned fields cover solid and surface volume, weight, support, static friction force, density, and friction coefficient. These outputs test whether spatial support remains stable while the predicted physical quantity changes.

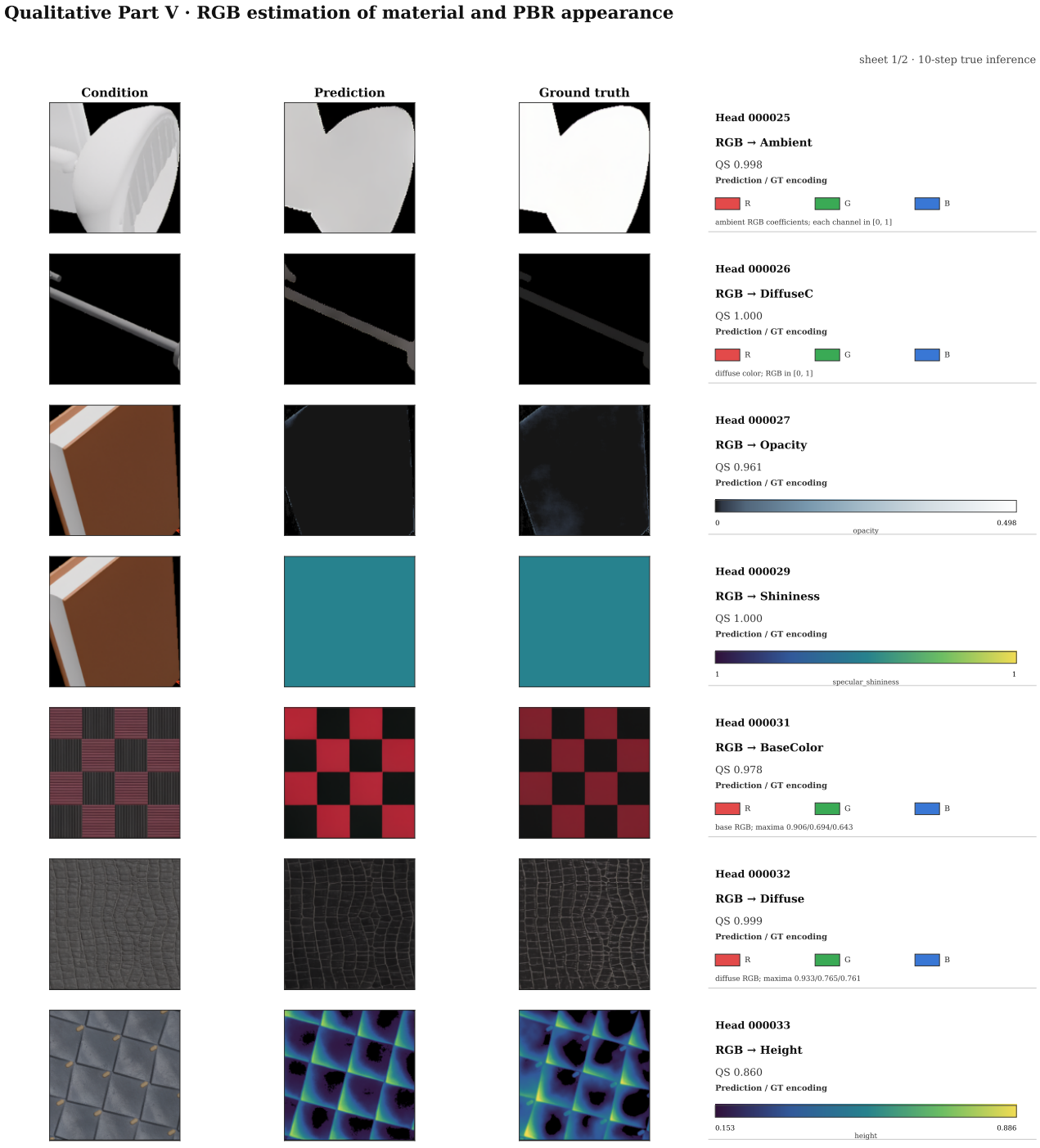


Figure 24. **Qualitative Part V: RGB estimation of material and PBR appearance (1/2).** Ambient and diffuse color, opacity, specular color, shininess, base color, and diffuse response separate spatial material support from the value assigned within it.

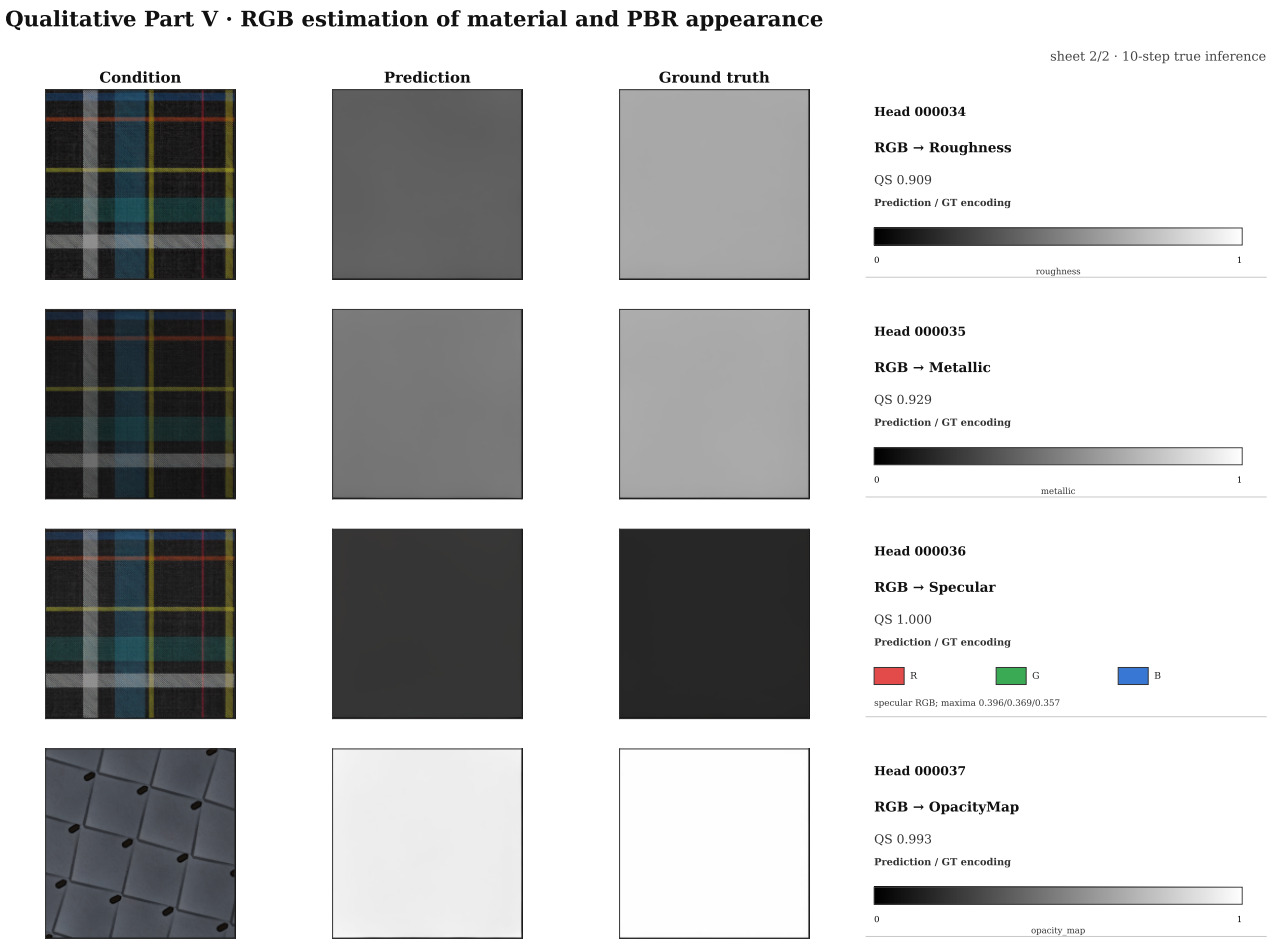


Figure 25. **Qualitative Part V: RGB estimation of material and PBR appearance (2/2).** Height, roughness, metallicity, specular response, and opacity maps complete the PBR interface. Prediction and ground truth use the same channel-specific encoding in every row.

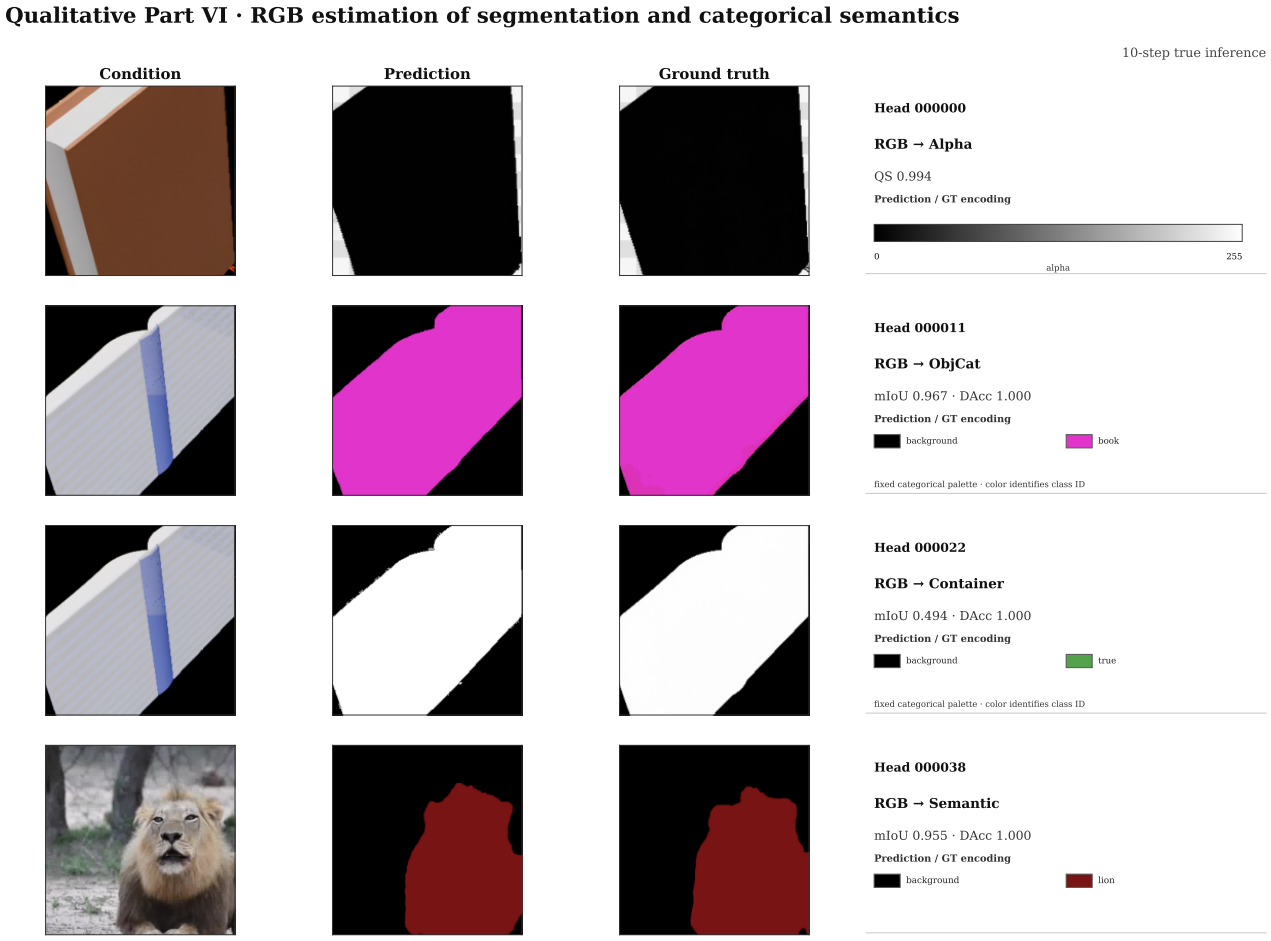


Figure 26. **Qualitative Part VI: RGB estimation of segmentation and semantics.** Alpha support, object category, container, and multi-class semantic masks isolate the spatial and categorical aspects of segmentation. The categorical rows show selected successful cases; aggregate accuracy remains reported by the quantitative matrices.

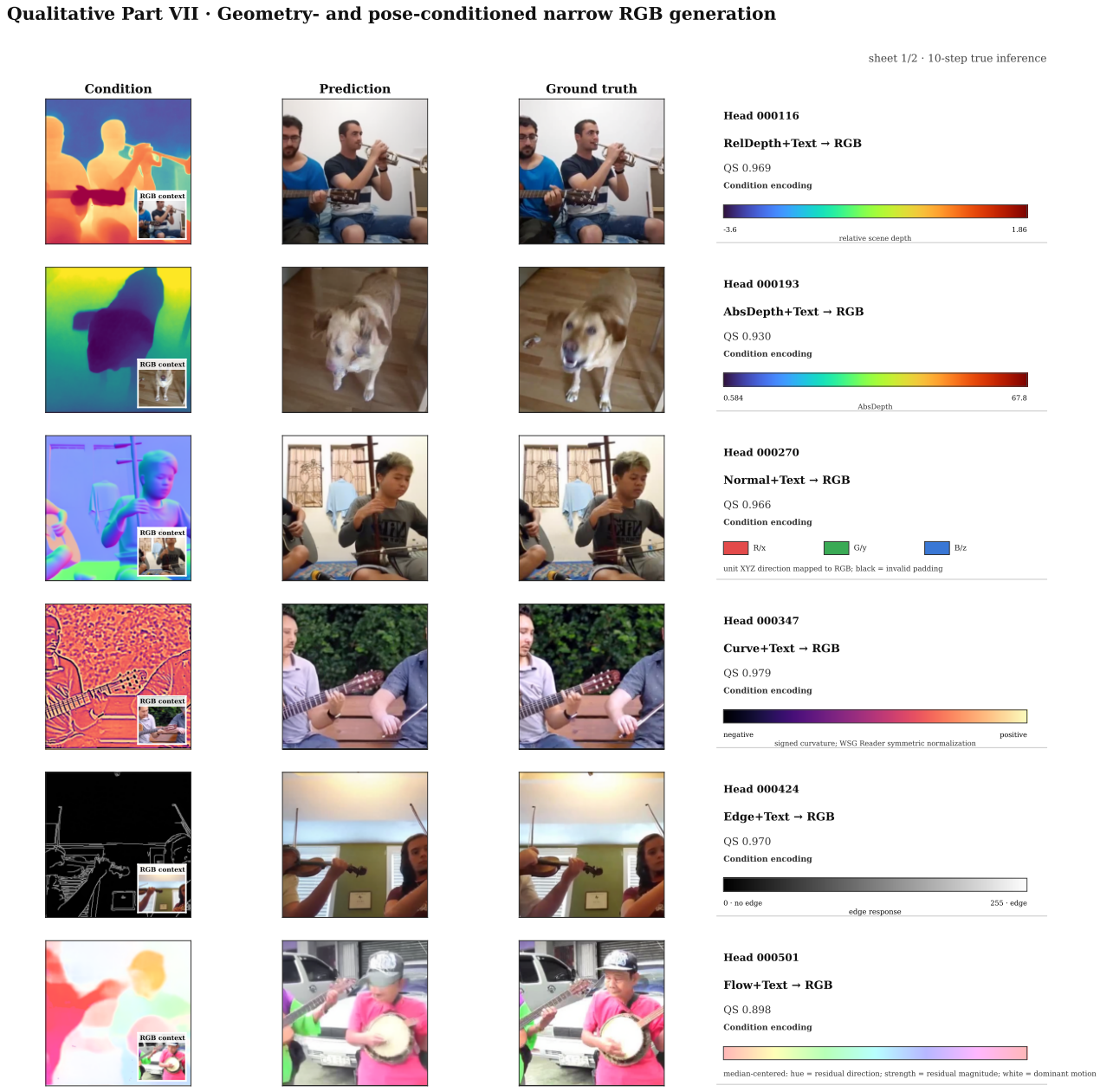


Figure 27. **Qualitative Part VII: geometry-conditioned narrow RGB generation (1/2).** Relative depth, absolute depth, normals, curvature, and edges supply the target-frame structure. Historical RGB context is embedded inside the condition panel and never substitutes for the evaluation target frame.

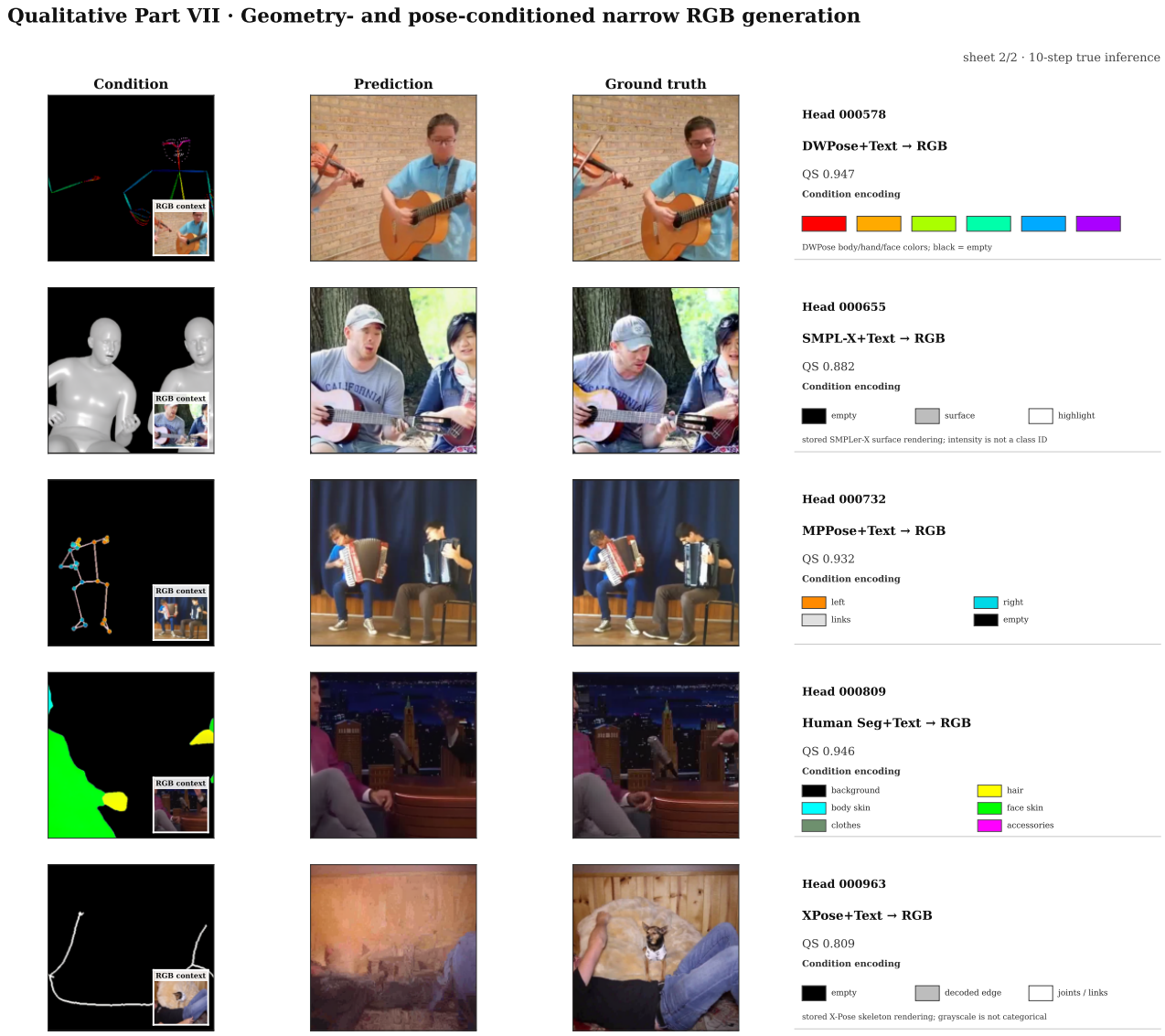


Figure 28. **Qualitative Part VII: motion- and pose-conditioned narrow RGB generation (2/2).** DWPose, SMPL-X, MediaPipe pose and MediaPipe Seg, and XPose test whether sparse, articulated, or semantic pose conditions redirect the generated frame while text and the historical reference preserve scene identity.

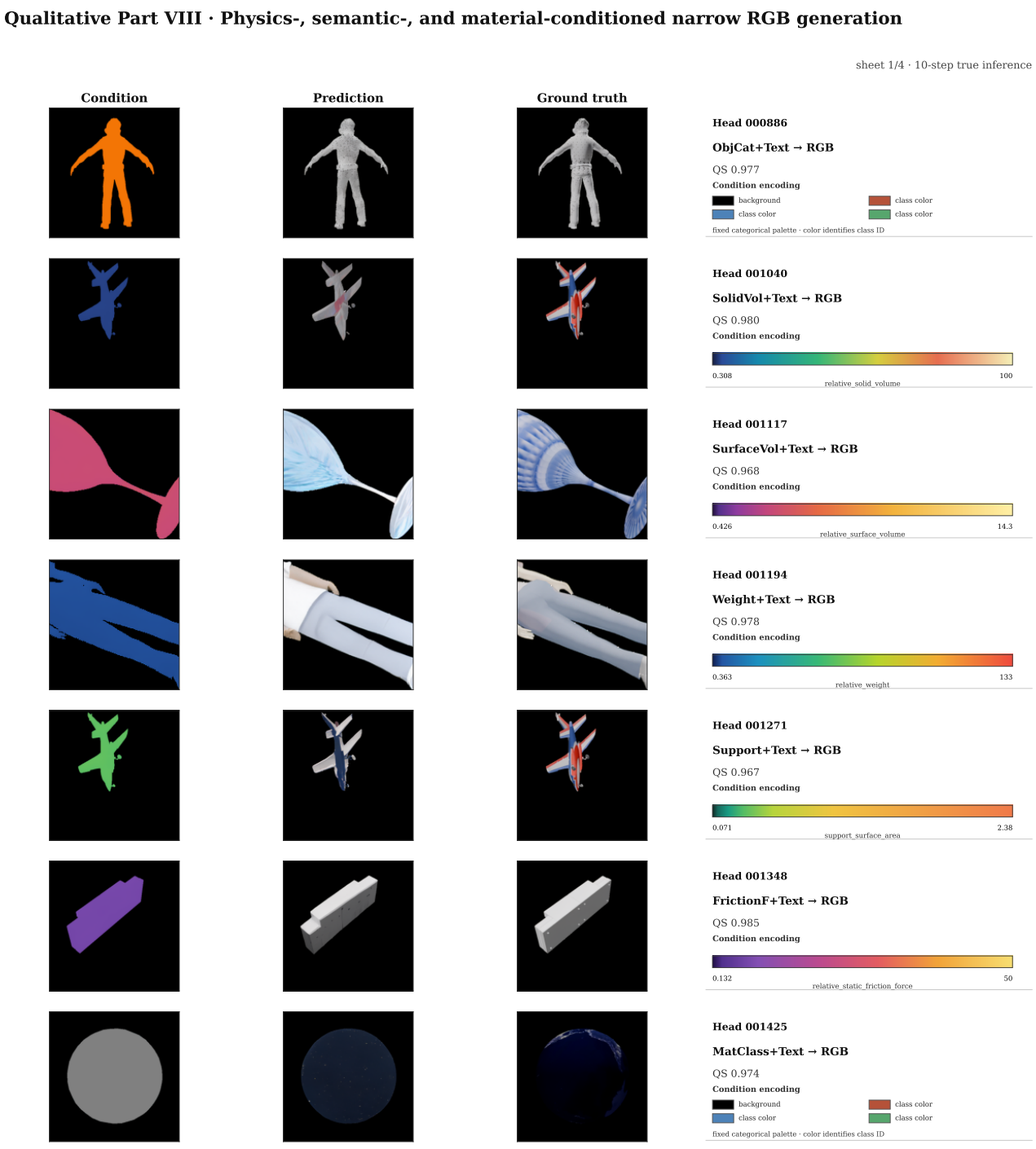


Figure 29. **Qualitative Part VIII: physics-, semantic-, and material-conditioned narrow RGB generation (1/4).** Object class, volume, weight, support, and friction fields begin the exhaustive non-geometric source sweep.

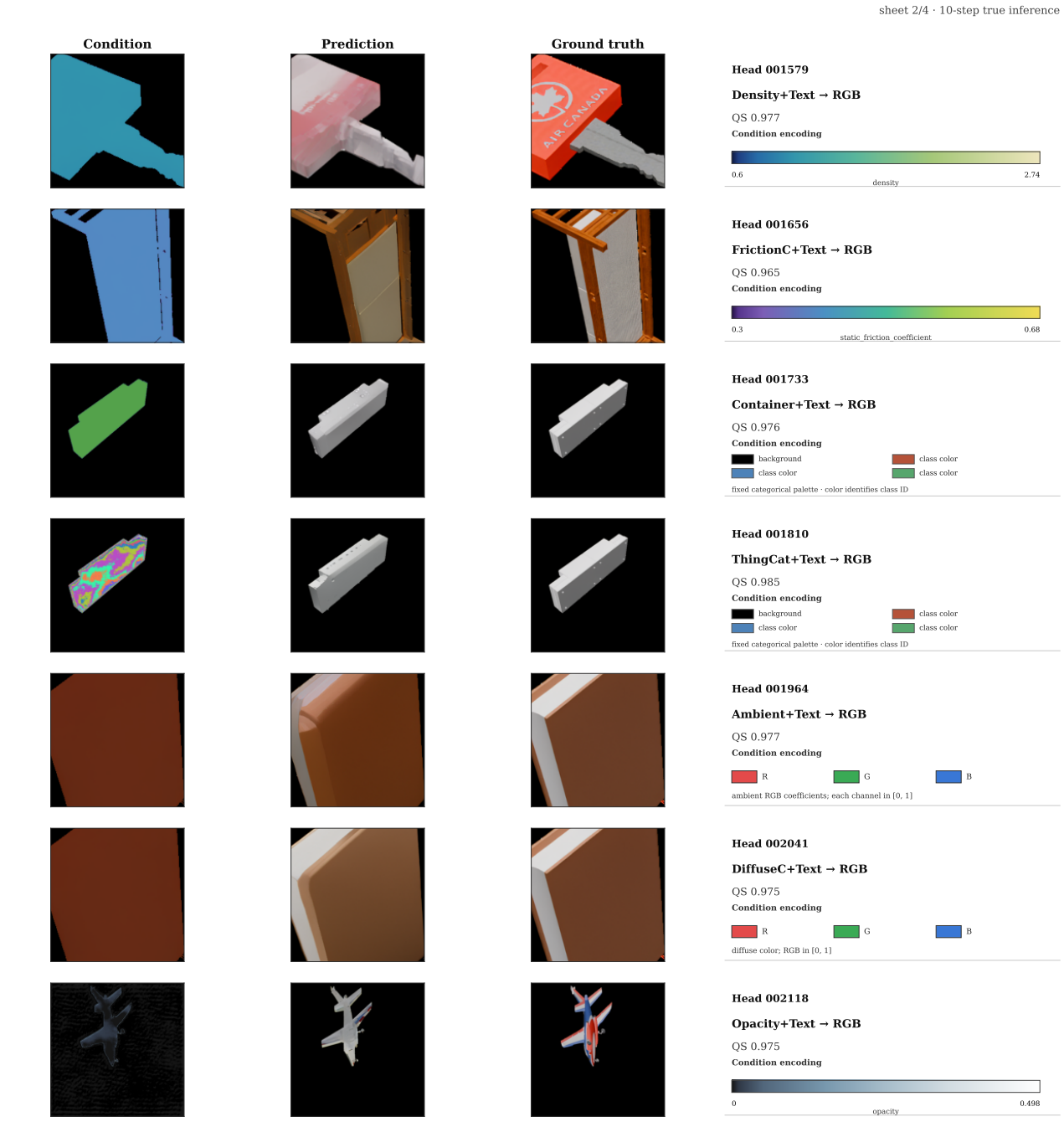


Figure 30. **Qualitative Part VIII: physics-, semantic-, and material-conditioned narrow RGB generation (2/4).** Material class, density, friction coefficient, container, entity category, ambient color, and diffuse color test controls whose information is concentrated inside object support rather than distributed across the image.

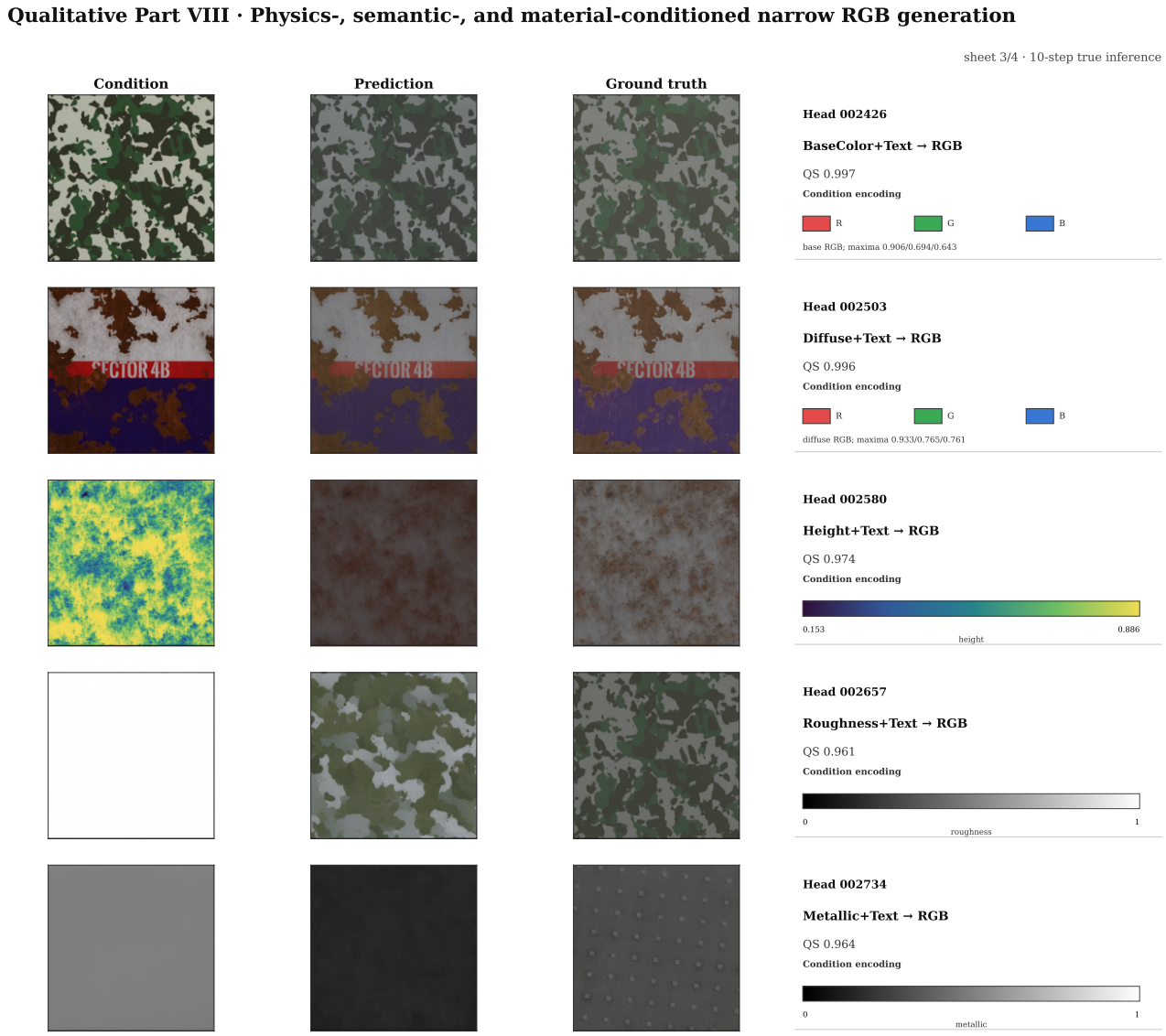


Figure 31. **Qualitative Part VIII: physics-, semantic-, and material-conditioned narrow RGB generation (3/4).** Base color, diffuse response, height, roughness, and metallicity show how appearance controls interact with text and RGB context.

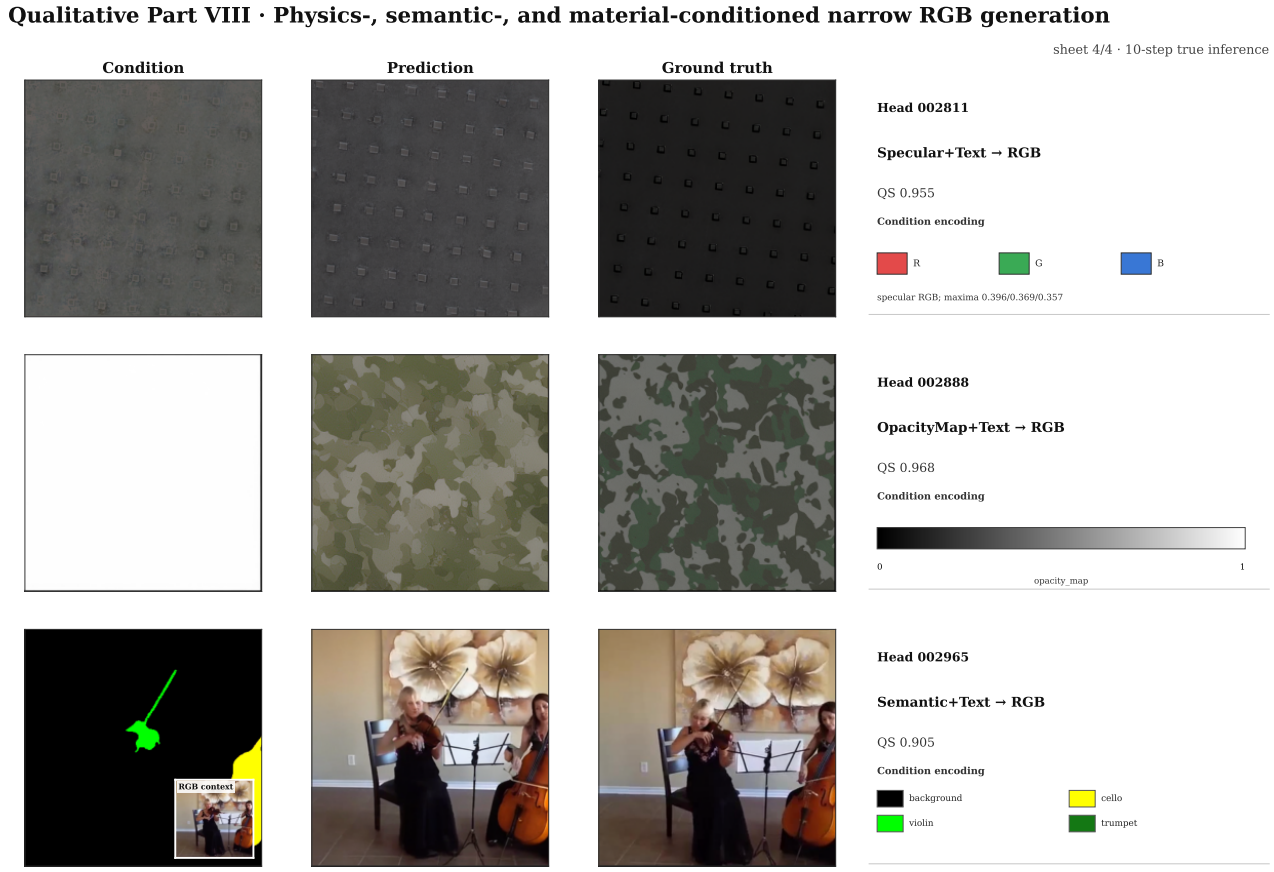


Figure 32. **Qualitative Part VIII: physics-, semantic-, and material-conditioned narrow RGB generation (4/4).** Specular response, opacity maps, and semantic masks complete the 33 visually readable narrow-generation source modalities.

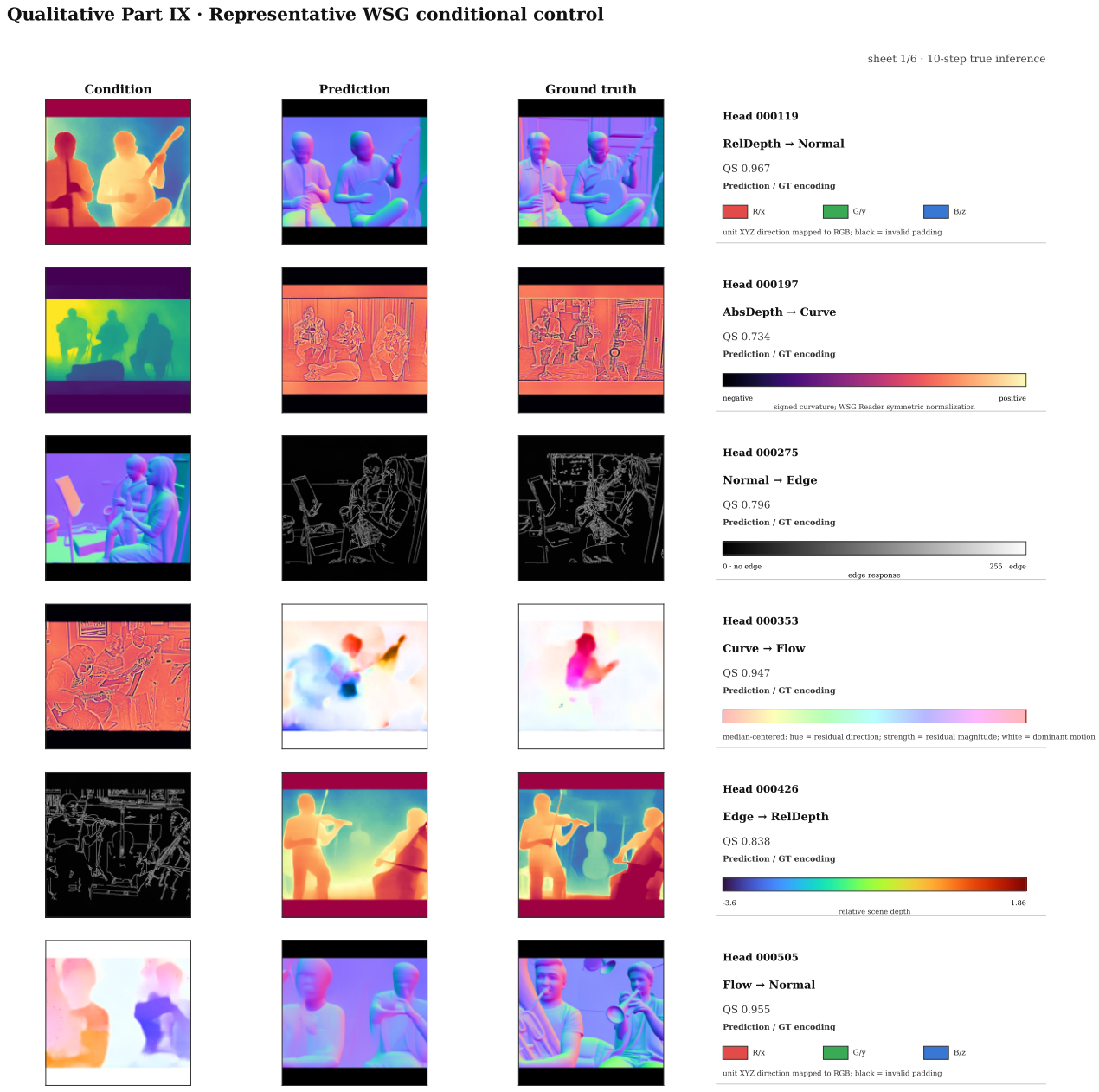


Figure 33. **Qualitative Part IX: representative WSG conditional control (1/6).** Six geometry-to-geometry transfers cover depth, normals, curvature, edges, and flow without auxiliary RGB evidence.

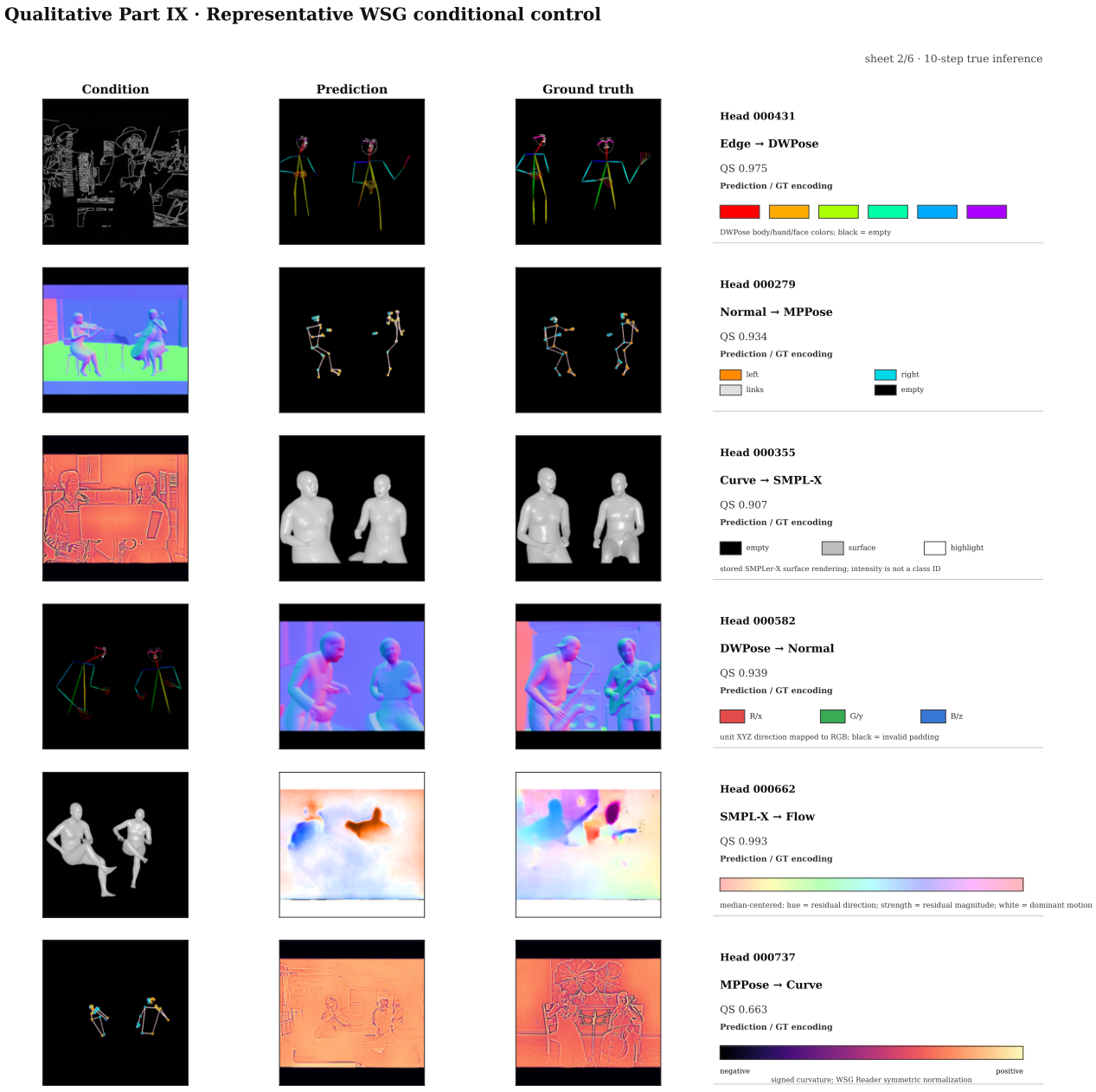


Figure 34. **Qualitative Part IX: representative WSG conditional control (2/6).** Geometry-to-pose and pose-to-geometry mappings test whether dense surface structure and sparse articulated structure can supervise one another.

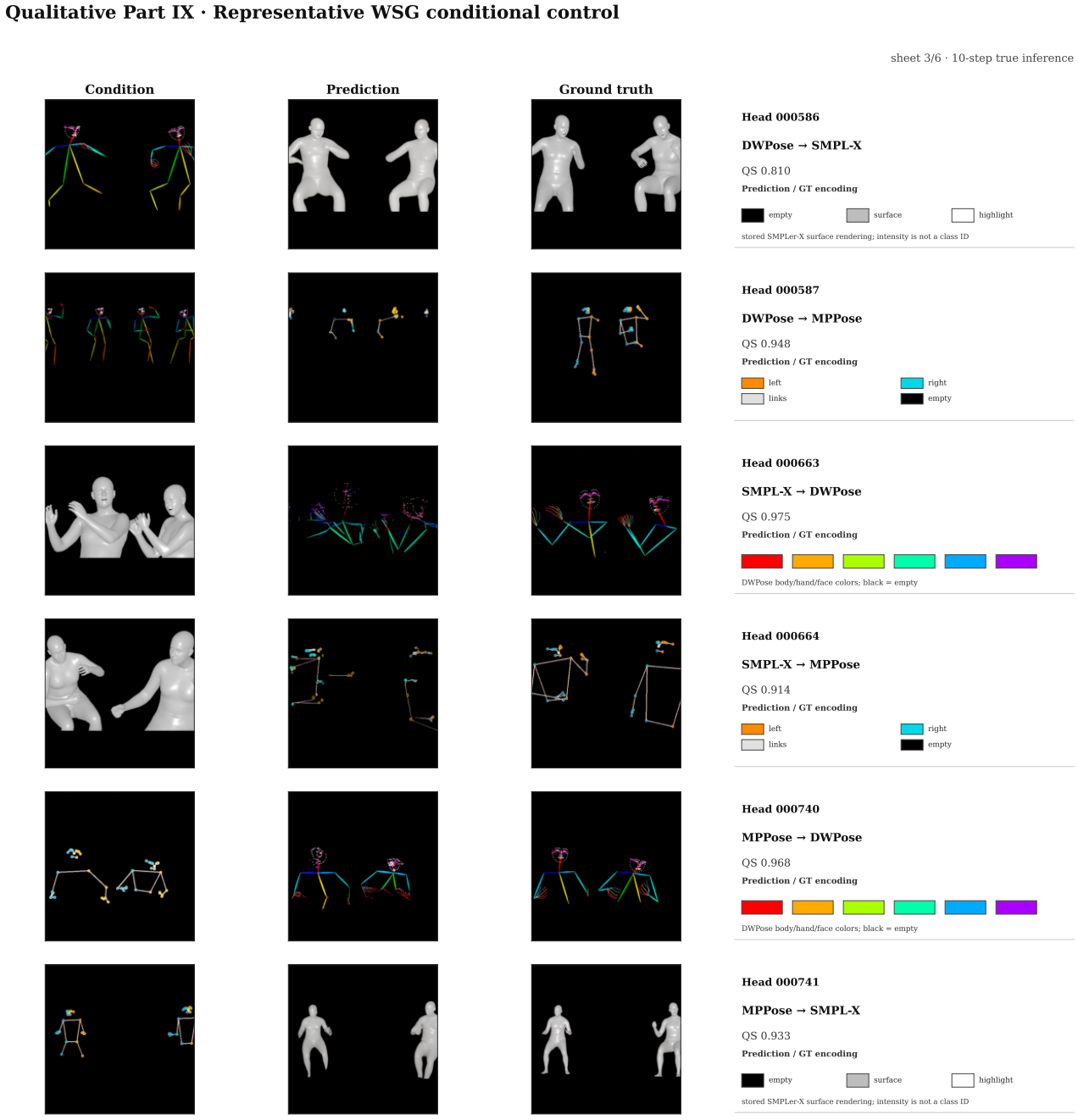


Figure 35. **Qualitative Part IX: representative WSG conditional control (3/6).** All six directed conversions among DWPose, SMPL-X, and MediaPipe pose are shown, exposing the common articulated structure beneath distinct encodings.

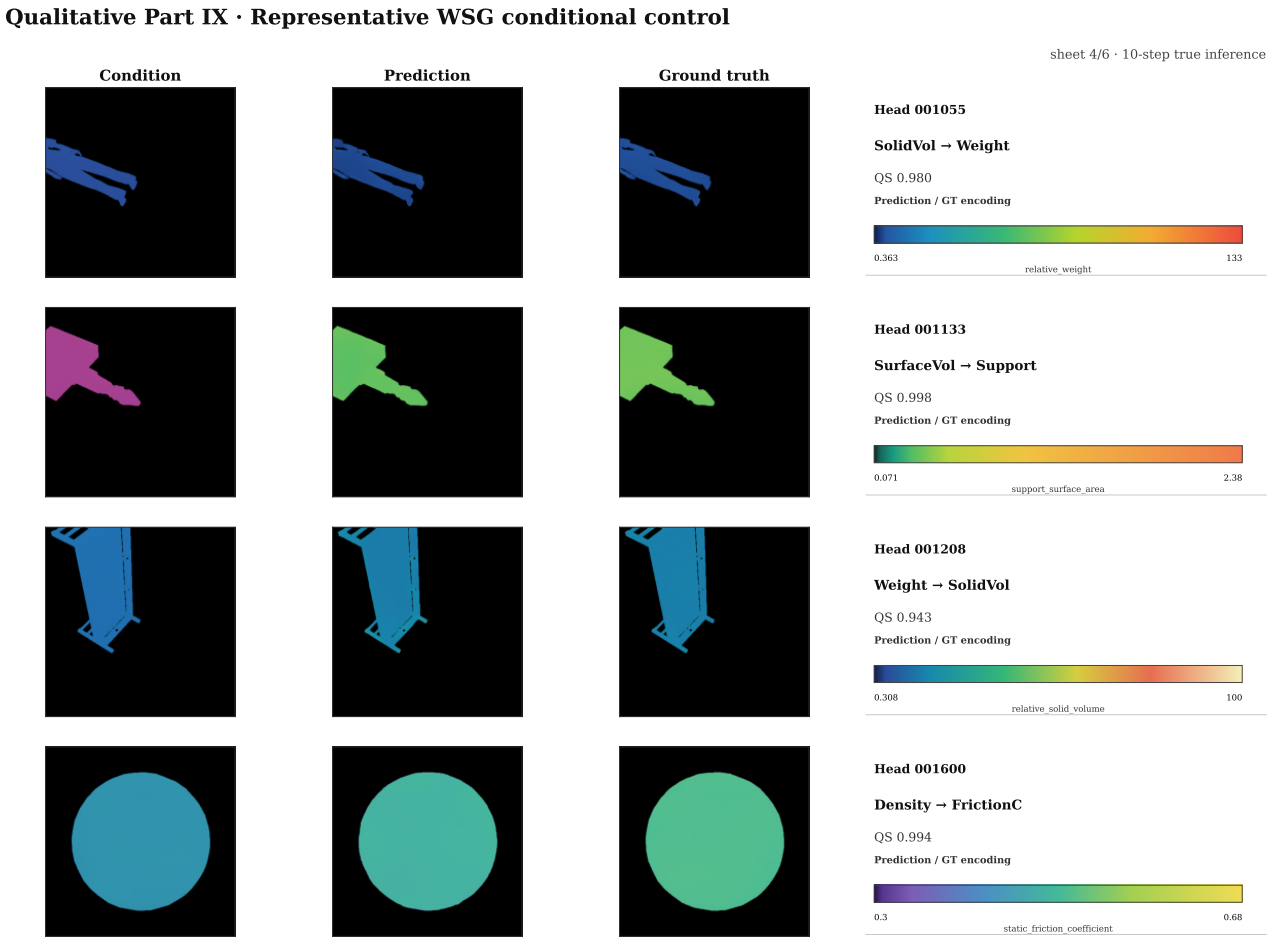


Figure 36. **Qualitative Part IX: representative WSG conditional control (4/6).** Four physical-property transfers probe volume, weight, support, density, and friction while preserving object-aligned spatial support.

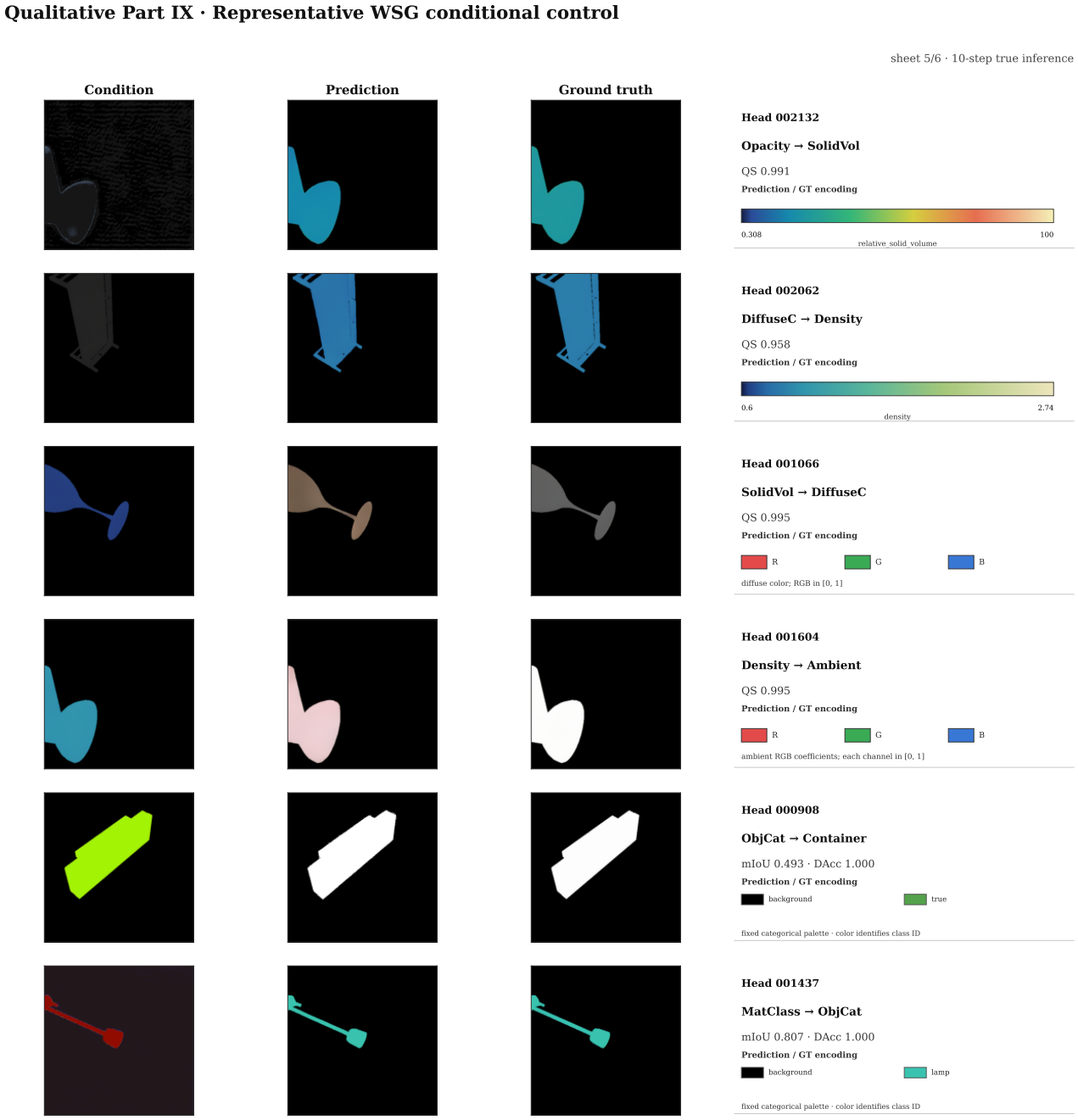


Figure 37. **Qualitative Part IX: representative WSG conditional control (5/6).** Four material-to-physics and physics-to-material mappings test whether shared object support is sufficient for translation between distinct value spaces; two categorical rows test label-aware transport.

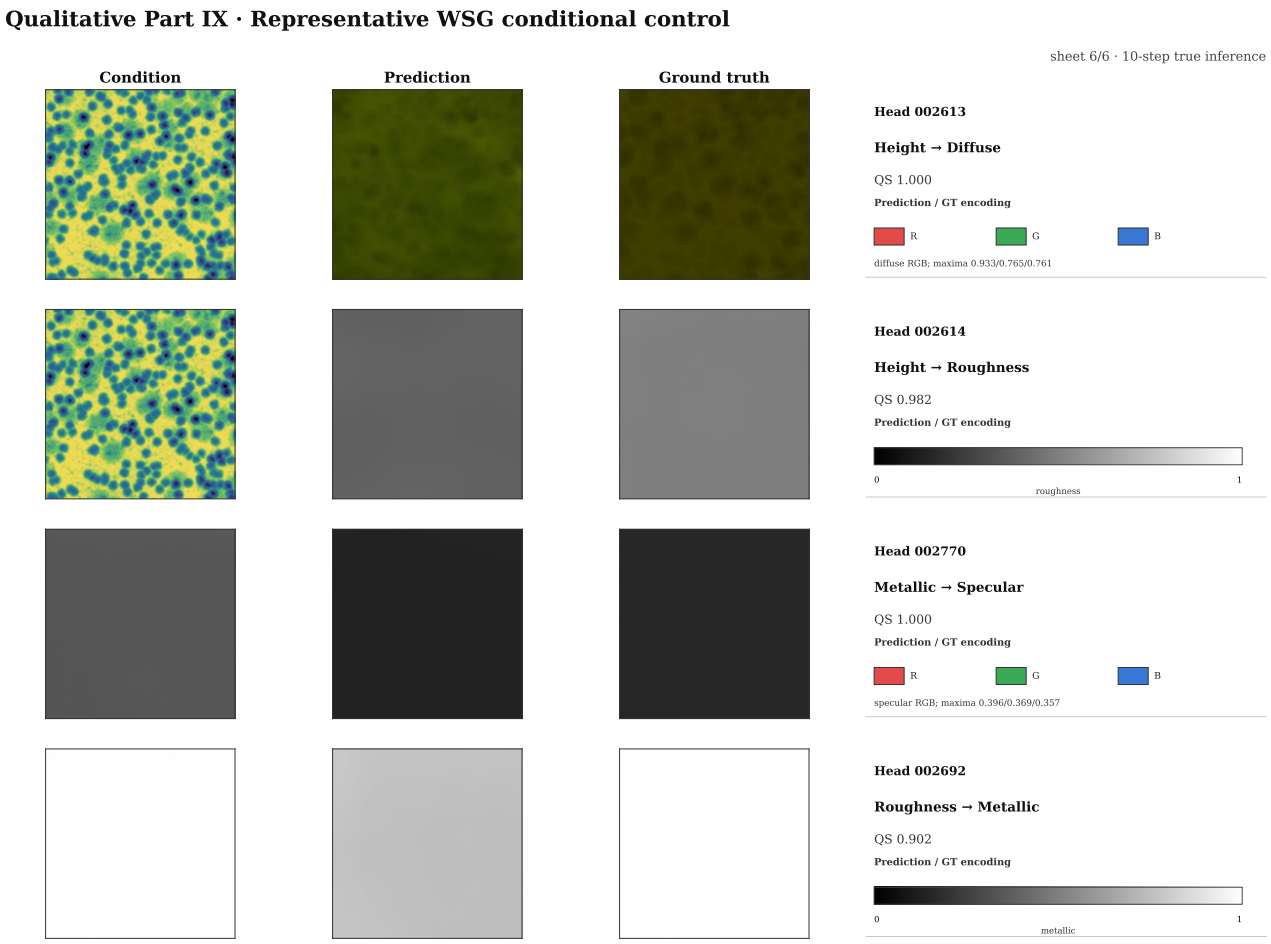


Figure 38. **Qualitative Part IX: representative WSG conditional control (6/6).** Four PBR conversions isolate translation among base color, diffuse, roughness, metallicity, and specular response.

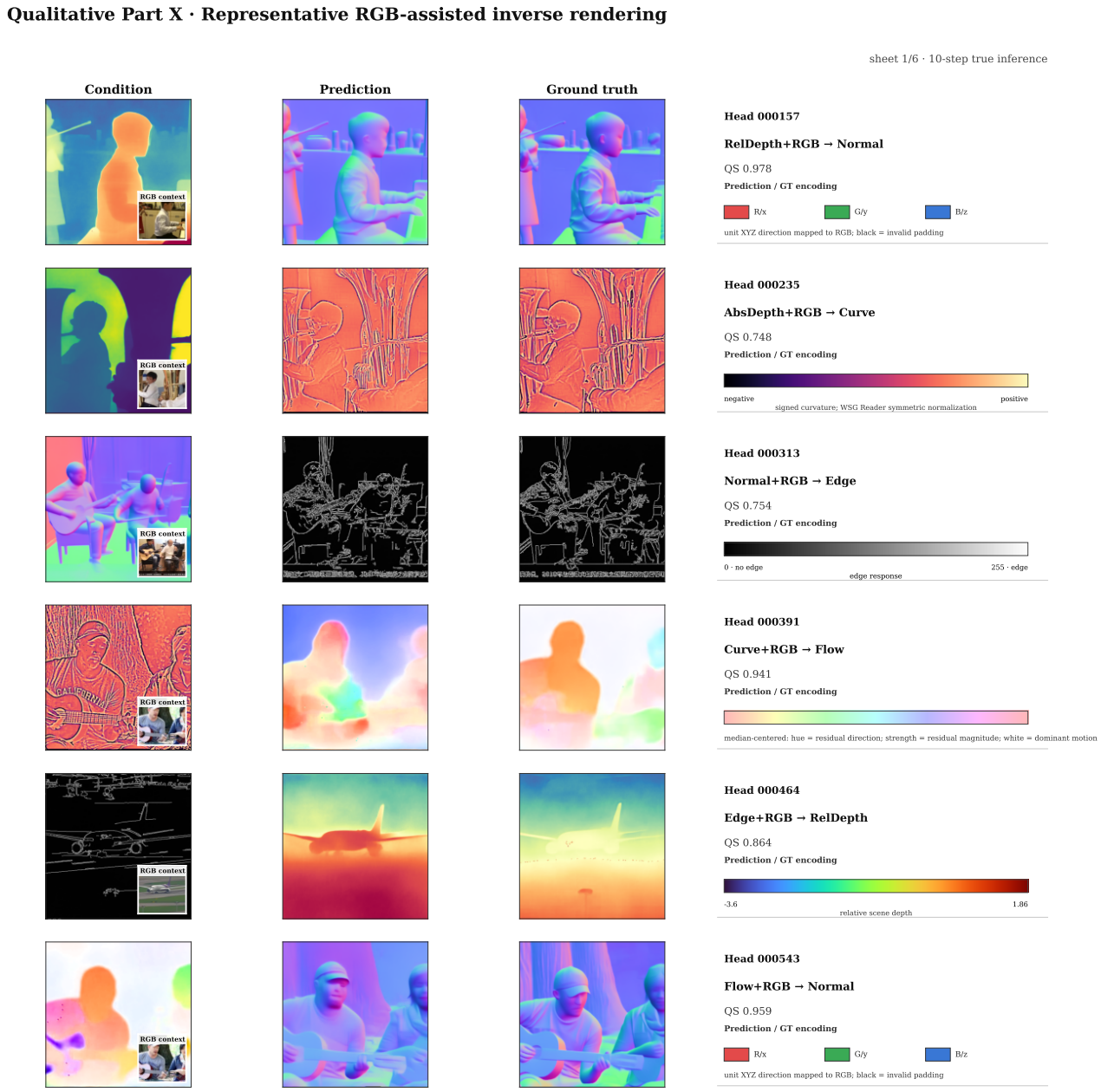


Figure 39. **Qualitative Part X: representative RGB-assisted inverse rendering (1/6).** Geometry-to-geometry transfers repeat Part IX with observed RGB evidence, making scene-specific boundaries and identity available to the Head.

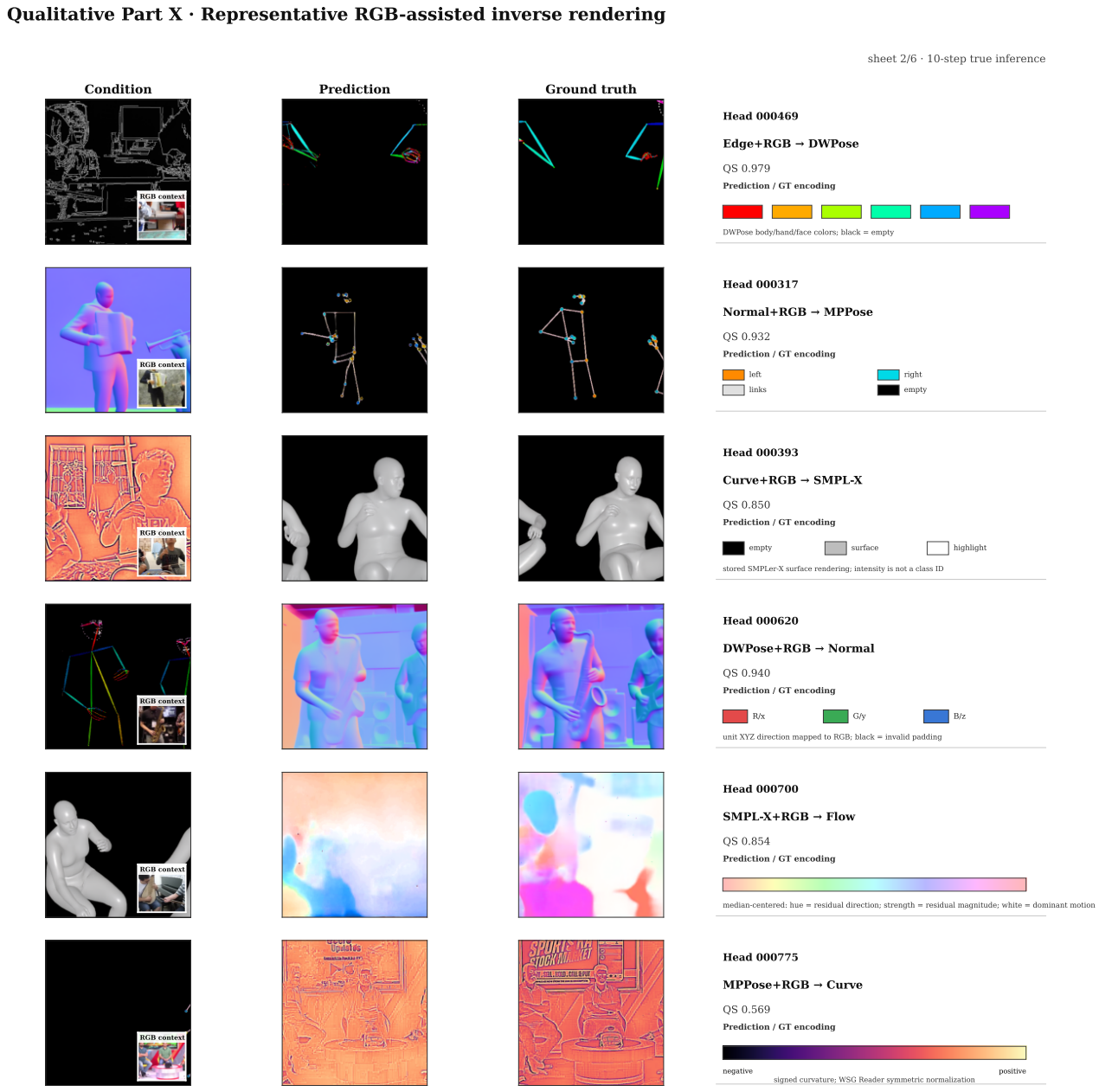


Figure 40. **Qualitative Part X: representative RGB-assisted inverse rendering (2/6).** Geometry-to-pose and pose-to-geometry mappings show how RGB evidence resolves structure that is underdetermined by the first condition alone.

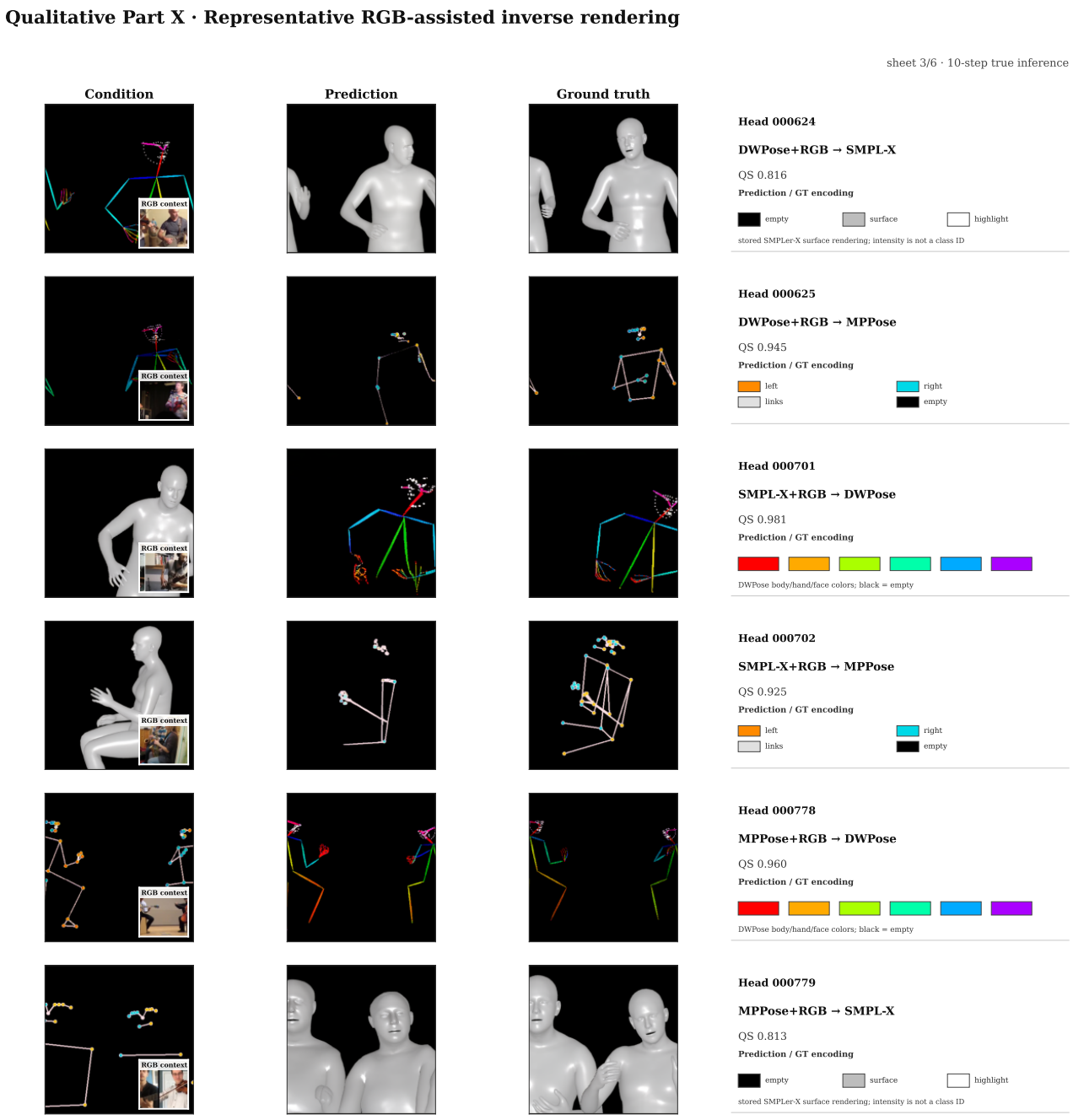


Figure 41. **Qualitative Part X: representative RGB-assisted inverse rendering (3/6).** Six directed pose conversions retain their explicit condition while using the RGB observation to recover scene-specific body extent and visibility.

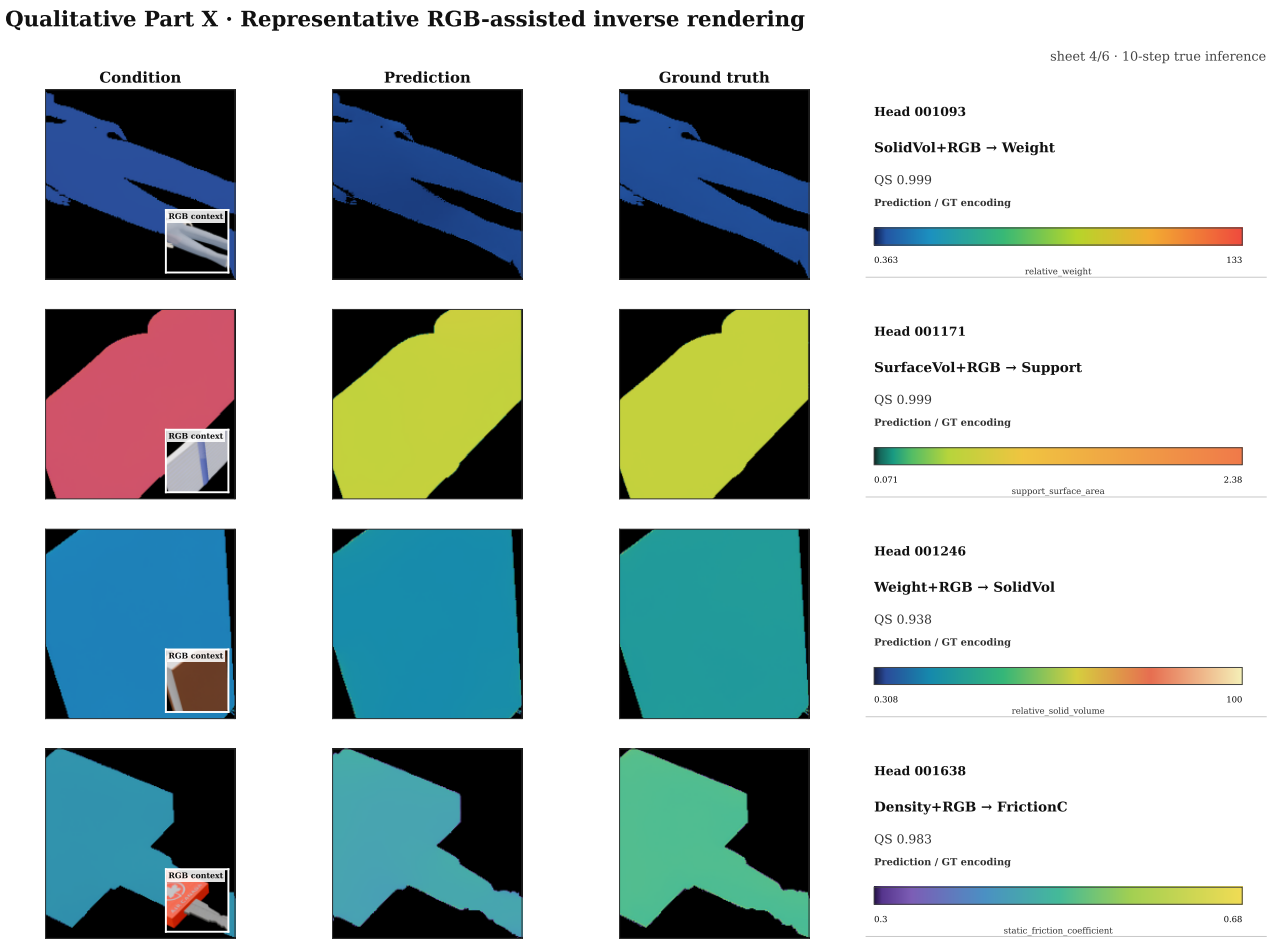


Figure 42. **Qualitative Part X: representative RGB-assisted inverse rendering (4/6).** Physical-property mappings combine object-aligned structured evidence with RGB boundaries to stabilize support and predicted value.

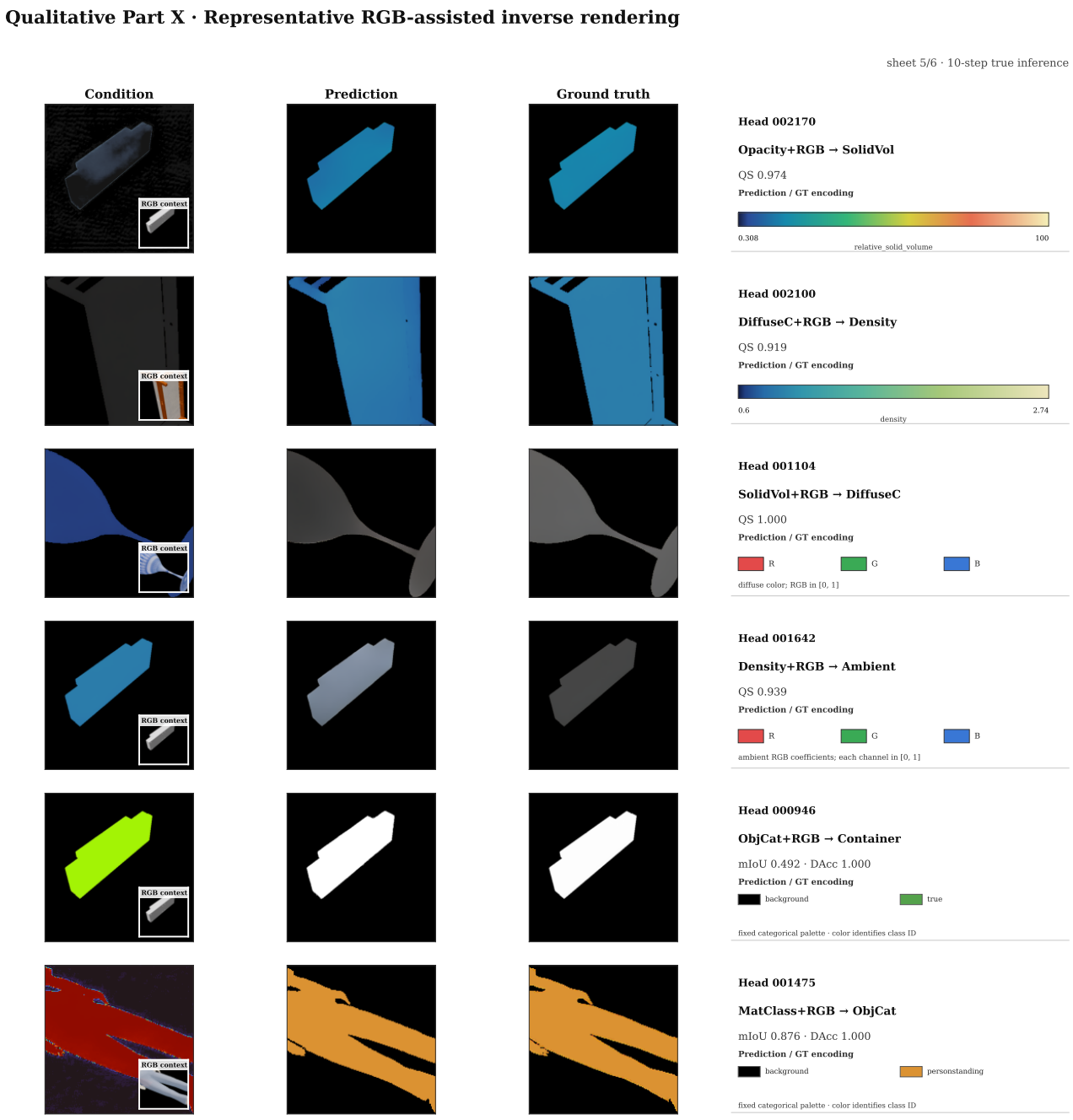


Figure 43. **Qualitative Part X: representative RGB-assisted inverse rendering (5/6).** Four cross-family transfers show where visible object identity complements the structured input; two categorical rows retain label-aware evaluation.

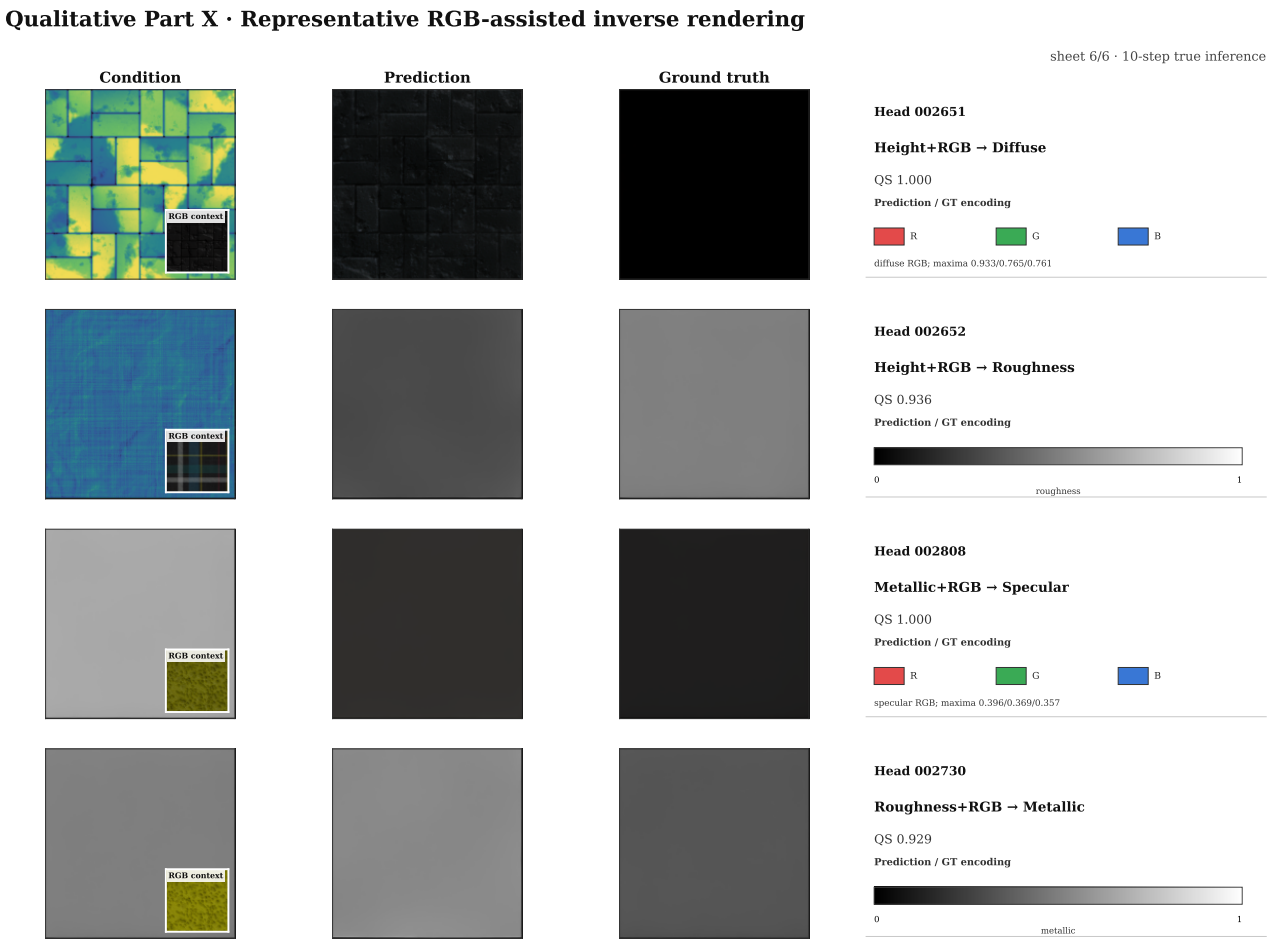


Figure 44. **Qualitative Part X: representative RGB-assisted inverse rendering (6/6).** Four PBR conversions complete the 32-mapping RGB-assisted slice of the 536-Head family.

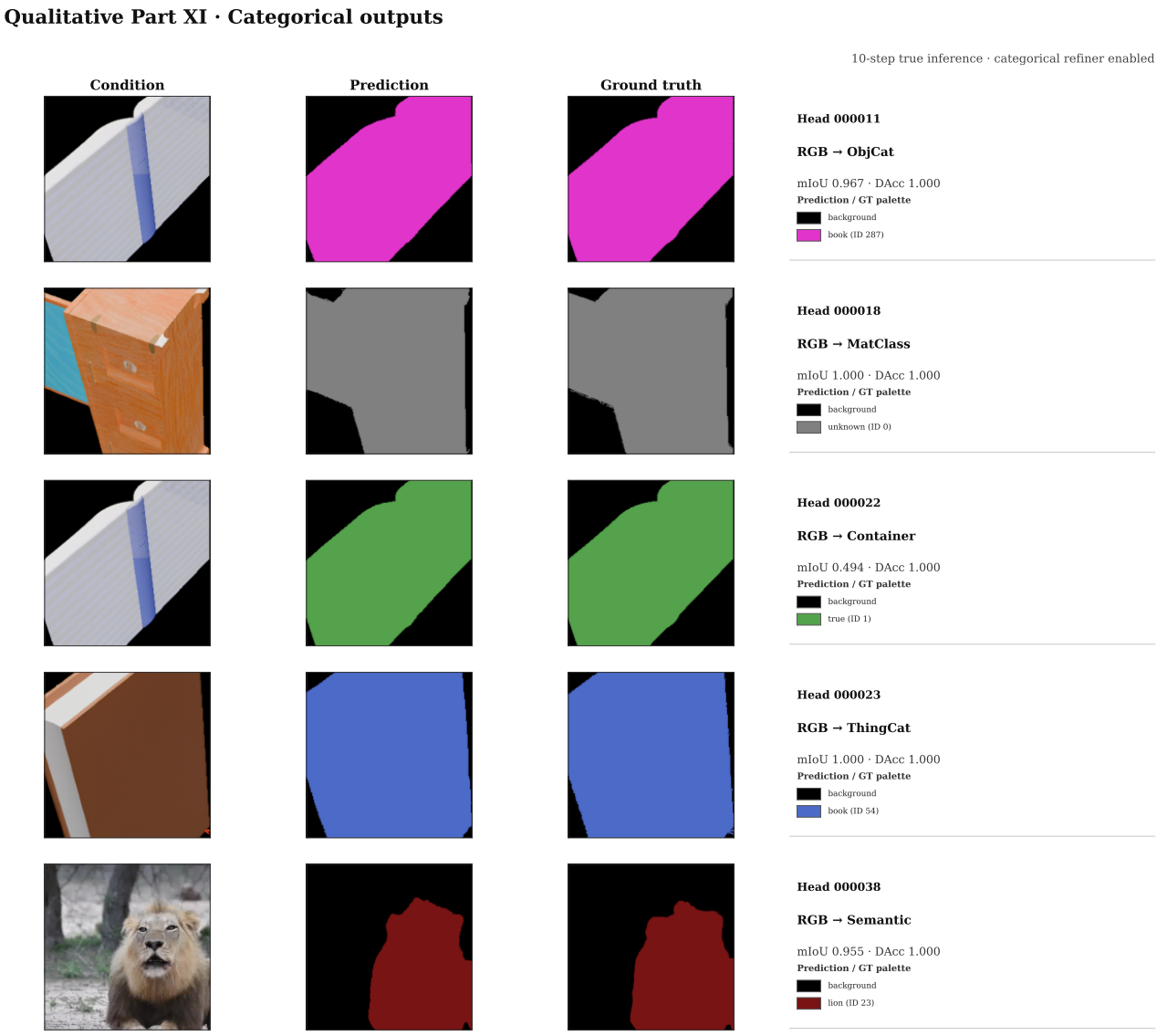


Figure 45. **Qualitative Part XI: categorical outputs.** One verified example is retained for each of the five categorical targets. Each row shows the condition, the final output of the complete categorical Head, and the corresponding ground truth.

6.4. Interface Comparison

Part I exposes an evidence asymmetry that is consistent with human geometric reasoning. In the Normal-to-Edge row of Figure 20, the Head produces a visually cleaner boundary map than the pseudo-ground-truth rendering: smooth surface normals retain geometric discontinuities while suppressing texture- and illumination-specific high-frequency variation. The objective is therefore not to reproduce every incidental detail of the source representation, but to transport the evidence needed by the requested target. Conversely, when a target depends on boundaries or appearance not determined by a structured condition alone, RGB-assisted inverse rendering restores scene-specific cues from the observation. Omitting RGB yields a geometry-supported abstraction; supplying RGB restores observation-specific detail. We next examine every remaining qualitative figure in order, separating spatial fidelity, value calibration, and categorical assignment.

6.5. RGB Estimation

Geometry and motion. Figure 21 shows that the same RGB condition can be translated into six geometrically distinct representations. Normals, curvature, and absolute depth are especially stable (QS 0.959, 0.896, and 0.936). Relative depth and the direction-consistent flow example retain the dominant scene layout (0.850 and 0.968), while Edge is less uniform (0.837) because thin boundaries are only weakly determined by local appearance. The contrast is consistent with human geometric reasoning—surface orientation can remain clear even when an exact metric distance or boundary convention is ambiguous.

Articulated structure. Figure 22 tests DWPose, SMPL-X, MediaPipe Pose, XPose, MediaPipe Seg, and Universal Pose rather than six renderings of one dense field. Their common success is structural: joints remain connected in a recognizable body layout. In the third row, the prediction occasionally completes joints that are missing or malformed in the automatically produced MediaPipe pseudo ground truth. This does not make the proxy annotation infallible; instead, it shows that the learned human prior can denoise imperfect supervision. The fourth-row XPose targets are likewise produced by an automatic detector rather than manual annotation, and their quality remains a limitation of the present dataset. Nevertheless, their scale supplies enough weak supervision for the Head to recover characteristic skeletal structure. The result is evidence that the proposed paradigm can discover useful structure from large, imperfectly annotated corpora, while also identifying where better labels would improve the model. The MediaPipe Seg row complements these skeletal encodings with a dense estimate of human support. Its official six-region palette distinguishes background, hair, body skin, face skin, clothes, and accessories. We render the Head's generated field through this discrete palette before categorical-refiner correction, so the figure exposes both region support and direct class assignment. The final row extends the same image-to-structure formulation beyond fixed human representations: a custom Head recovers the category-independent red-joint, white-link Universal Pose used by MP-100 from an RGB observation.

Physical properties. Figure 23 contains seven object-aligned fields whose QS values lie between 0.974 and 1.000. Solid and surface volume, weight, support, friction, and density preserve nearly the same object support while changing the value carried inside it. This result should not be read as direct measurement of an object's physical constants; rather, it shows that the Head learns the dataset's encoded physical estimate without losing spatial correspondence to the observed object.

Material and PBR appearance I. Figure 24 compares ambient and diffuse color, opacity, specular color, shininess, base color, and diffuse response. The strong opacity, specular-color, shininess, and base-color examples show that broad, piecewise-smooth appearance attributes can be transported with little spatial drift. Ambient and diffuse color remain more sensitive to illumination, making them less directly recoverable than support or intrinsic appearance.

Material and PBR appearance II. Figure 25 exposes the harder part of the PBR interface. Height, metallicity, and specular response remain strong, whereas roughness and opacity maps deviate more visibly from the target. These channels require calibrated local intensity rather than only a correct object mask; consequently, a prediction may retain the right region while assigning an imperfect material value.

Segmentation and semantics. Figure 26 demonstrates that continuous Alpha and categorical masks can share the same generative interface. The selected material, entity, object, container, and semantic examples have high mIoU and dominant-class accuracy, so both the generated support and principal palette label are correct in these cases. They are deliberately readable successful examples, not a substitute for the aggregate matrices: Figure 45 later presents the final outputs of the complete categorical Heads.

6.6. Narrow RGB Generation

Geometric control. Figure 27 reverses the RGB-estimation direction. Depth, normals, curvature, edges, and flow now constrain RGB synthesis while text and optional historical RGB provide appearance. The six examples range

from QS 0.898 to 0.979; curvature and normal control are particularly stable, while flow is weakest. This ordering reflects how directly a condition fixes the target frame: static geometry constrains most pixels, whereas flow must be interpreted jointly with temporal and appearance context.

Human control. Figure 28 replaces dense geometry with articulated human representations. DWPose, SMPL-X, MediaPipe Pose/Class, and XPose all produce coherent RGB frames, with QS spanning 0.809–0.947. The lower SMPL-X and XPose scores expose the difficulty of reconstructing texture and background from sparse pose evidence. QS measures target-frame reconstruction and therefore complements, rather than replaces, visual inspection of whether the requested pose actually redirects the person.

Object-level physical control. Figure 29 covers object category, volume, weight, support, and friction conditions. Despite their low visual bandwidth, these fields yield recognizable frames because the physical and category channels narrow the set of compatible appearances.

Class and material control. Figure 30 continues with material class, density, friction coefficient, container, entity category, and ambient color. These conditions remain useful even when their nonzero information occupies only a small part of the canvas. Text and historical RGB remain necessary because the structured channel determines what must be preserved but does not fully specify photometric appearance.

Appearance control. Figure 31 separates direct color fields from derived PBR controls. Base color and diffuse response retain broad object support, while height, roughness, and metallicity constrain appearance more indirectly. The comparison shows how text and historical RGB resolve the many RGB images compatible with a low-bandwidth material field.

PBR and semantic control. Figure 32 completes the readable source sweep with specular response, opacity map, and semantic masks. The PBR fields preserve continuous appearance structure; a semantic mask constrains identity and support but intentionally leaves within-class texture and illumination underdetermined.

6.7. Conditional Control

Geometry-to-geometry transport. Figure 33 spans QS 0.734–0.967, revealing that geometric channels are not interchangeable merely because they share image coordinates. Relative-depth-to-Normal and Flow-to-Normal are strong, whereas Normal-to-Edge is lower because smooth normals suppress texture boundaries present in the target edge convention. The cleaner predicted edge is thus a meaningful geometric abstraction, even when it does not minimize every target pixel error.

Geometry and pose. Figure 34 crosses dense surfaces with sparse body structure. Edge-to-DWPose and DWPose-to-Normal reach 0.975 and 0.939, showing that visible boundaries can localize a body and a skeleton can guide coarse surface orientation. Curve-to-SMPL-X and MediaPipe-to-Curvature reach 0.907 and 0.663, as they require recovering quantities absent from the source rather than translating a shared structure.

Pose-system conversion. Figure 35 isolates six directed conversions among DWPose, SMPL-X, and MediaPipe Pose. All six retain strong joint organization, while DWPose-to-SMPL-X is the hardest at 0.810. The asymmetry is important: two pose systems can describe the same person yet expose different joints, surface priors, and occlusion conventions, so reversing an edge is not the same learning problem.

Physical-property transport. Figure 36 is close to a ceiling for volume, weight, support, density, and friction mappings (0.943–0.998). These tasks share object-aligned support and often vary mainly in the value assigned inside that support. The result explains the coherent high-score block in the quantitative matrix without implying that arbitrary physical variables are causally equivalent.

Cross-family transport. Figure 37 tests deliberately less homogeneous pairs. Opacity-to-SolidVolume and DiffuseColor-to-Density remain strong, while Density-to-Ambient is harder. Spatial support transfers readily across families, but an appearance value cannot always be inferred from a physical scalar. The two categorical rows additionally test whether shared support can carry a discrete label rather than only a continuous value.

PBR transport. Figure 38 isolates four PBR conversions. Base-color and diffuse relations remain strongly constrained by shared material support. The fourth row, roughness-to-metallic, exposes a fundamentally weaker form of evidence: roughness does not determine metallicity, so the Head can only exploit learned visual regularities—for example, that metallic surfaces are often smooth—to form a plausible estimate. This is not deterministic recovery, and the resulting degradation is expected. Nevertheless, the prediction goes beyond copying aligned pixels: it reveals a cross-attribute visual association that can be strengthened as supervision and coverage improve.

6.8. RGB-Assisted Inverse Rendering

Geometry with observed appearance. Figure 39 is the RGB-assisted counterpart of Figure 33. RGB improves Edge-to-RelativeDepth (0.838 to 0.864) and modestly improves both Normal outputs, but Normal-to-Edge remains difficult and decreases from 0.796 to 0.754 in this sample. The reference therefore contributes missing scene evidence; it does not guarantee a monotonic gain when the target convention conflicts with the structural condition.

Geometry–pose inversion. Figure 40 shows the same mixed but interpretable behavior. RGB yields 0.932 for Normal-to-MediaPipe and 0.850 for Curvature-to-SMPL-X, while MediaPipe-to-Curvature remains the weakest example at 0.569. Visible appearance resolves body extent and occlusion, but a single RGB frame still cannot uniquely determine every differential surface quantity.

Pose inversion. Figure 41 provides the cleanest evidence that the reference image complements, rather than replaces, the structured condition. Compared directly with the condition-only conversions in Figure 35, the RGB-assisted results are visibly more complete and stable: the observation resolves person count, visible body extent, and occlusion cues that a sparse pose representation leaves ambiguous. Three of the six directed conversions improve quantitatively; DWPose-to-SMPL-X rises from 0.810 to 0.816 and SMPL-X-to-MediaPipe from 0.914 to 0.925. The pose channel specifies the requested representation, while RGB restores the scene evidence needed to instantiate it accurately.

Physical inversion. Figure 42 retains the near-ceiling behavior of the physical block, now with RGB evidence. All four displayed mappings remain between 0.938 and 0.999 QS. Differences are small because object support was already available in the structured condition; RGB mainly sharpens visible boundaries.

Cross-family inversion. Figure 43 shows where observation is most useful. Density-to-Ambient improves substantially, while opacity-, diffuse-, and solid-volume-conditioned mappings remain stable. The categorical rows show the same principle for discrete targets: RGB can clarify visible object identity, but cannot repair a missing spatial support region. The evidence supports selective assistance rather than a claim that RGB universally dominates structured transport.

PBR inversion. Figure 44 isolates four RGB-assisted PBR conversions. The largest gain occurs for roughness-to-metallic, where visible appearance supplies evidence absent from roughness alone. Directly related color and specular fields change less because their structured conditions already carry most of the required support and value information. This contrast localizes when observation repairs a cross-modal ambiguity.

6.9. Categorical Output Analysis

For categorical targets, the categorical refiner is an integral component of the corresponding Head and is enabled by default during inference. Figure 45 presents the resulting end-to-end outputs for Semantic, ThingCat, Container, MatClass, and ObjCat. The shared palettes make both spatial extent and class identity directly comparable with the ground truth.

6.10. Sparse Digitization and Extended Tasks

WSM can admit named nodes outside the fixed 256-channel WSG schema, provided that the new variable has an aligned, machine-readable representation. This gives sparse annotations and newly digitized structures the same operational status as an existing image-like endpoint without renumbering the released dense graph or retraining the shared generator.

Universal Pose is the first explicit example of this extension path. We render the category-independent MP-100 landmark annotations as sparse red joints and white links on a black canvas, register RGB and Universal Pose as named nodes in a custom task graph, and train one additional Channel Head. The final row of Figure 22 is a 10-step generation result on the evaluation set from that Head. It extends the fixed human-pose representations above it to the articulated structure of animals and objects while preserving the same condition–prediction–target interface.

To keep quantitative claims stable as the system grows, extended tasks remain separate from the frozen 1,148-task aggregate until they pass the same evaluation protocol. Subsequent additions can be reported in this subsection with their data source, task definition, selected model, metrics, and qualitative artifact. The task bank can therefore grow continually while each released result remains reproducible and verifiable.

7. Discussion

7.1. Experimental Evidence

What our three experimental sections establish. Our ablation study shows how TunnelingNet acquires a new visual route. We find that attention topology, evidence routing, tunnel placement, and high-noise trajectory supervision all shape how quickly a task-specific Head learns its target representation while preserving the RGB prior. The resulting configuration produces recognizable structure within tens of updates and reaches up to approximately $100\times$ greater training efficiency than the parameter-matched native-attention baseline. Our quantitative study then asks how widely this recipe holds: it evaluates 1,148 visual tasks, with one independently deployable Channel Head per task, across 39 output groups, four interfaces, and 18,191 end-to-end inference evaluations. Our qualitative study lets us inspect the visual relationships behind those numbers across geometry, articulated humans, motion, physical properties, materials, PBR appearance, segmentation, categorical identity, and RGB synthesis. Together, our experiments give a consistent picture: a shared generator can learn a wide vocabulary of World Snapshot transformations when each route receives focused parameters and clearly routed evidence.

Scale and strength of our baseline. We summarize continuous outputs with the higher-is-better quality score QS and categorical outputs with dominant-class accuracy (DAcc), abbreviated together as QS/DAcc. Across the 1,147 tasks with a valid primary score, 779 (67.9%) exceed $\mathrm{QS/DAcc} = 0.8$. Within the 969 continuous tasks, 744 (76.8%) exceed $\mathrm{QS} = 0.8$, 537 (55.4%) exceed $\mathrm{QS} = 0.9$, and the mean QS is 0.860. Strong results appear across all four interfaces: 37/38 narrow-generation tasks, 367/449 RGB-assisted inverse-rendering tasks, 312/449 condition-only control tasks, and 28/33 continuous estimation or reconstruction tasks exceed 0.8. A conservative sensitivity check that removes the two near-constant output groups flagged by the artifact analysis still gives 666/891 continuous tasks (74.7%) above 0.8. These measurements give us confidence that the strong region extends across hundreds of heterogeneous mappings. The paired matrices also show a natural dependence on the visual evidence shared by the condition and target.

7.2. Unified Task Graph and Model Scale

Why the directed graph matters. Each WSM task is a directed path from available evidence to a World Snapshot target. Moving a variable from the condition side to the target side asks the model to generate that variable. RGB therefore becomes one node in a much larger generative domain that also includes geometry, pose, motion, material, physics, semantics, and audio. This construction makes training naturally parallel: every edge shares the frozen backbone, while its small tunnel can be scheduled, optimized, and validated independently. It also makes capabilities compositional. A generated variable can become evidence for the next edge, so longer visual chains can be assembled from learned paths.

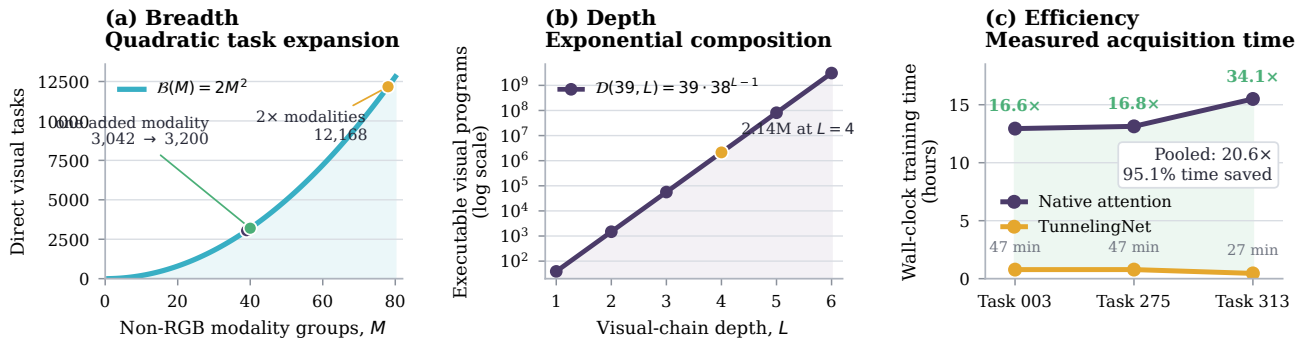
The complete modular WSM comprises one shared backbone, a resident collection of trained Heads and an explicit task router. Independent Head optimization and physical sharding of the Head files enable parallel training and modular distribution without changing this model boundary. All task Heads remain registered together, while a task key sparsely activates one edge and a visual chain activates a selected sequence. This resident-versus-active organization is analogous to sparsely activated expert models. WSM routes with an explicit visual-task key; conventional expert models commonly learn their gating. We verified that each of the 1,148 frozen dense task Heads contains exactly 12,510,208 tunnel parameters, giving 14.362B task-Head parameters in total. Together with the 3.9B visual backbone and 0.282B parameters from the 180 deployed categorical refiners, the complete resident visual parameter bank is approximately 18.54B parameters. In this complete deployment form, WSM v1 Augustus is an 18B-class model with 1,148 selectable visual tasks.[7] A single inference route activates one 12.51M tunnel, so its active visual parameter count remains approximately 3.913B; categorical routes add only 1.54–1.60M active refiner parameters. Replicating the 3.9B backbone once for every evaluated task would already require approximately 4.48T visual parameters. Sharing one backbone therefore reduces this task-library parameterization by about $241\times$, while retaining a dedicated generative route for every task.

7.3. Mathematical Scalability of the Unified Law

Toward a law-scale foundation for vision. Most computer-vision models are built around one task, and broadly capable systems usually collect a small set of task-specific interfaces. WSM v1 places 1,148 evaluated tasks inside one generative system, while its released dense task graph configures 3,042 possible mappings. Relative to the conventional scope of one to ten tasks, this expands the task count by roughly two to three orders of magnitude. It also changes the unit of progress: we can grow one directed generative graph by adding World Snapshot variables and routes, and every new route becomes part of the same model.

The central result therefore reaches beyond any individual benchmark score. A scientific law is useful when it compresses many apparently separate phenomena into one practical rule. Our experiments provide large-scale

[7] This count excludes the shared text encoder and VAE.



Unified-Law scaling: expand the modality vocabulary, compose deeper visual routes, and train each route in parallel.

Figure 46. **Mathematical scalability of the Unified Law.** (a) The four-interface task graph grows as $\mathcal{B}(M) = 2M^2$; the current 39 non-RGB modality groups define 3,042 dense mappings. (b) Composing directed routes produces exponentially more executable visual programs as chain depth increases. (c) Directly logged wall-clock measurements on three matched tasks show 16.6–34.1× faster route acquisition with TunnelingNet, with a pooled 20.6× speedup and 95.1% time reduction. The three panels measure complementary properties: the width of the visual-task graph, the depth of reasoning over that graph, and the cost of instantiating its learned routes.

evidence for such a compression in computer vision: geometry, pose, motion, material, physics, semantics, audio, and appearance become World Snapshots, and generation becomes their common computational operation. TunnelingNet makes this principle trainable in parallel, the directed graph makes it compositional, and the shared backbone makes thousands of routes deployable as one WSM. This is the law-scale ambition of the Unified Law for Visual Tasks: to organize a growing field of visual capabilities through one reusable generative principle. WSM v1 Augustus is its first large-scale realization, and the 3,042 mappings in its dense task graph provide an immediate path for expansion.

The scalability implied by the Unified Law admits a compact mathematical form. We characterize it along three measurable axes: the breadth of directly trained tasks, the depth of composed visual programs, and the cost of acquiring the learned routes. Together, these expressions formalize how one unified generative principle expands into a growing visual-task system.

Unified-Law scalability formula: breadth and depth. Let M be the number of non-RGB World Snapshot modality groups. The four-interface dense graph contains

$$\mathcal{B}(M) \equiv T_{\text{direct}}(M) = M + M + M(M-1) + M(M-1) = 2M^2, \tag{14}$$

covering RGB estimation, narrow RGB generation, directed conditional control, and RGB-assisted inverse rendering. The current $M = 39$ configuration gives exactly 3,042 dense mappings. Once a new modality is aligned frame-by-frame with the existing World Snapshots, it can appear as condition and target throughout the graph. For K added modalities, the direct expansion is

$$\Delta_K T = T(M+K) - T(M) = 4MK + 2K^2. \tag{15}$$

At the present scale, one new modality creates 158 direct tasks and ten create 1,760. Doubling the modality vocabulary from 39 to 78 quadruples the dense task graph from 3,042 to 12,168 mappings.

Composition grows even faster. Starting from RGB, the condition-only graph already admits

$$\mathcal{D}(M, L) \equiv P_L = M(M-1)^{L-1} \tag{16}$$

visual chains of length L. For $M = 39$, lengths one through four yield 39, 1,482, 56,316, and 2,140,008 possible directed paths. Equations 14–16 form the breadth–depth expansion rule induced by the Unified Law. The breadth $\mathcal{B}(M)$ grows quadratically with the modality vocabulary, while the reasoning depth $\mathcal{D}(M, L)$ grows exponentially with visual-chain length. These axes reinforce each other: frame-aligned modalities raise the branching base available at every reasoning step, and longer visual chains raise its exponent. WSM can therefore expand the variety of visual questions and the depth of the programs used to answer them at the same time.

The third scaling axis is the cost of acquiring these routes. Let t_i denote the measured wall-clock training time of task route i and let G be the number of parallel workers. Because every task Head shares a frozen backbone and can be optimized independently, the wall-clock cost of constructing an N-task bank is approximately

$$\mathcal{C}_{\text{train}}(N, G) \simeq \frac{1}{G}\sum_{i=1}^{N} t_i \simeq \frac{N\bar{t}}{G}. \tag{17}$$

This expression makes the practical value of the graph explicit: capability breadth can grow across tasks while acquisition time decreases almost linearly with available parallelism. In our controlled wall-clock records for Tasks 003, 275, and 313, native attention required 12h56m32s, 13h08m09s, and 15h29m53s, whereas TunnelingNet required 46m55s, 47m01s, and 27m18s at the selected milestones. The corresponding speedups are $16.6\times$, $16.8\times$, and $34.1\times$. Pooled across the three tasks, training falls from 41.58 to 2.02 GPU-hours, a measured $20.6\times$ acceleration and a 95.1% reduction in acquisition time. This wall-clock result complements the 19.0–$38.5\times$ update-count acceleration and the 77–$100\times$ earlier emergence of recognizable structure in Sec. 4.

7.4. Unifying and Expanding Computer Vision

This section tests the central scope claim of the Unified Law. We ask two questions. First, can the established tasks of computer vision be expressed within the WSM interface system? Second, after those established tasks are removed, how many tasks evaluated by WSM define genuinely new visual problems? The first question measures *unification*; the second measures *expansion*.

We use Papers with Code as the historical reference because it maintained one of the broadest public catalogs of computer-vision research, connecting named tasks with papers, datasets, code, and benchmarks. Its final public records contain 1,477 visual or vision-related tasks, which we use as a representative inventory of the established field [65]. For unification, we examine whether each historical task can be written as a route through the WSM interfaces. For expansion, we compare the 1,148 tasks evaluated by WSM v1 with this inventory and count the evaluated tasks that have no historical counterpart. The complete matching rules and task-level records are provided in Sec. 12.

The result is direct. All 1,477 historical visual tasks can be expressed within the WSM interfaces, giving 100% compatibility with the representative historical inventory. After historical overlaps are removed, 655–1,056 of the 1,148 evaluated WSM tasks are new. Relative to the 1,477 established visual tasks, WSM therefore expands the computer-vision task space by 44.3–71.5%. Figure 47 presents these two results.

Together, these findings establish both parts of the claim. WSM provides one generative interface system for the established field of computer vision, and it substantially enlarges that field with executable task definitions and trained baselines. Its dense graph already contains 3,042 mappings, and each newly aligned modality creates further families of estimation, generation, control, and reconstruction tasks. We call this direction *Task-Generative Computer Vision*: a visual framework that unifies the existing task space and systematically creates new tasks beyond it.

Position and interpretation of evidence. WSM makes three complementary claims, each supported at its appropriate experimental scale. First, WSG and the four interfaces provide one operational rule for expressing heterogeneous visual computations as directed generation. Second, WSM v1 turns that rule into an executable system: 1,148 distinct input–output tasks have their own trained TunnelingNets and end-to-end inference results within one shared-backbone model. Here, *unified* refers to the shared representation, generative backbone, route-acquisition rule and router. Independently optimized Heads are sparsely activated TunnelingNets within that model, not replications of the full backbone. Third, the historical analysis shows that the four interfaces cover 1,477 historical visual tasks, while the 1,148 figure separately measures tasks that were actually trained and evaluated. Keeping these quantities distinct makes the breadth claim directly auditable.

The validation scale was deliberately increased from tens of representative routes, to hundreds of cross-family mappings, and finally to 1,148 frozen tasks. Prior generalist vision systems commonly remain at tens of tasks, and reports beyond one hundred are rare. WSM therefore moves the empirical frontier to 1,148 separately trained and evaluated tasks: more than an order of magnitude beyond even a 100-task system, 29–$38\times$ beyond a 30–40-task system, and two to three orders beyond the conventional one-to-ten task setting. These levels answer different questions: matched MTA–NJA ablations isolate TunnelingNet acquisition behavior; QS/DAcc and 18,191 generated outputs measure system-wide coverage; task-native metrics establish target-specific quality; and AVIS, post-release imagery, and independent public annotations establish transfer after acquisition. In this last setting, *zero-shot* and *training-free* mean that a route is trained once and then applied to unseen datasets and annotation conventions without dataset-specific retraining. Likewise, the reported budget ratios measure marginal task acquisition after the respective pretrained priors, whereas the matched attention ablations isolate the contribution of TunnelingNet itself. This layered protocol evaluates scale without equating coverage with identical ceiling accuracy: the complete score distributions and disclosed failure regions identify which Heads merit further improvement while preserving the system-level evidence that one method can acquire, register, and execute more than one thousand visual tasks.

Finally, the breadth and chain equations describe the tasks available in the graph. An input–output mapping is a distinct task because changing either its available evidence or its target changes the learned conditional distribution; opposite directions are therefore different visual problems. The chain count measures the compositional paths made available by trained edges, while the directly

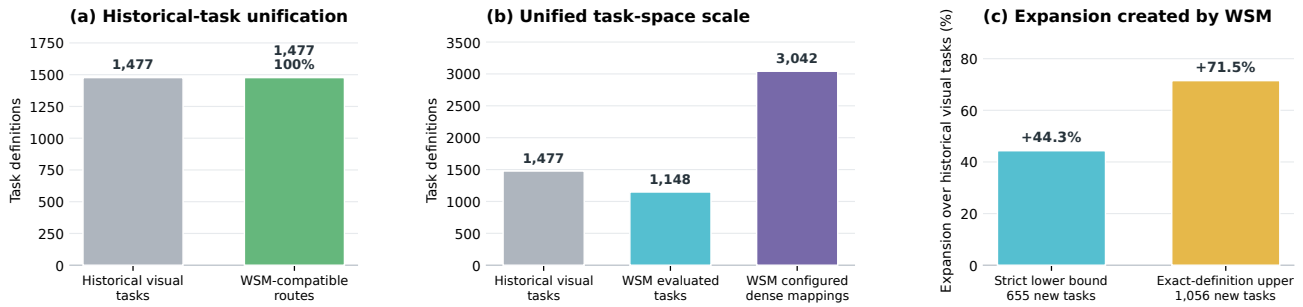


Figure 47. **WSM unifies and expands computer vision.** Papers with Code recorded 1,477 visual or vision-related tasks, providing a representative inventory of the existing field. (a) WSM can express all 1,477 tasks within its unified interfaces. (b) WSM v1 evaluates 1,148 tasks within a dense graph of 3,042 configured mappings. (c) After removing overlaps with existing tasks, 655–1,056 evaluated WSM tasks are new, expanding the established computer-vision task inventory by 44.3–71.5%. Detailed calculations are provided in Sec. 12.

reported claims remain anchored to trained TunnelingNets and observed outputs. Together, these definitions separate demonstrated execution, cross-dataset transfer, interface-level scope, and future composition without weakening the central result: WSM is the first realization of this principle at thousand-task scale.

7.5. Formal Scope and Falsifiability of the Unified Law

Why a law, and why vision now? Empirical laws in artificial intelligence capture reproducible regularities over defined regimes. The canonical study *Scaling Laws for Neural Language Models* [66] examines cross-entropy under one autoregressive language-modeling objective: the models are trained on WebText2, and the analysis establishes predictive power laws over model size, dataset size, and compute, with trends spanning six to eight orders of magnitude. Its importance does not come from separately validating every NLP task, but from discovering a stable empirical relation that compresses observations and predicts behavior outside individual runs. In task diversity, that evidence centers on one learning objective; WSM instead realizes one visual organizing principle through 1,148 separately trained visual tasks, 39 output groups, four interfaces, and 18,191 generated outputs. On a count of learning objectives, one is less than 0.1% of 1,148. The two laws are validated along different scientific axes: a scaling law is tested by depth along model, data, and compute scales, whereas a unifying law is tested by breadth across heterogeneous visual tasks. The breadth of 1,148 tasks directly supports the unification claim. The broader scientific precedent is clear: an empirical law in AI is established by a reproducible regularity over a defined regime.

The law and its empirical realization. The Unified Law has two deliberately separated levels. The law itself is a representation statement: once the information required by a visual task can cross the model boundary without losing a task-relevant distinction, the task can be written as conditional generation of requested world fields from known world fields. WSM and TunnelingNet provide an empirical realization of this statement: a shared generator can acquire many such directed relations through small, independently installed routes. The first level does not depend on a particular parameterization; the measured parameter counts and training times belong to the second.

Formally, let T map known fields C to requested fields Y under data distribution $\mathcal{D}_T$, let $s = (s_C, s_Y)$ be a fixed serialization, and write $y_a \sim_T y_b$ when two targets are equivalent under the task's native answer and evaluation. Define the *preservation condition* $\pi(T, s) \in \{0, 1\}$ by

$$\begin{aligned} \pi(T,s) = 1 \quad \Longleftrightarrow \quad & s_C(c_a) = s_C(c_b) \\ & \Rightarrow p_T(Y \mid c_a) = p_T(Y \mid c_b), \\ & s_Y(y_a) = s_Y(y_b) \\ & \Rightarrow y_a \sim_T y_b. \end{aligned} \tag{18}$$

Thus the condition code preserves every input distinction that can change the correct answer distribution, and the target code preserves every distinction that the task evaluates. For a model that observes only this fixed serialization, $\pi = 0$ is an information boundary: no later architecture, objective, or amount of training can recover a distinction that s has already erased. A different serialization may restore it; the impossibility claim is therefore about the declared s, not about the task in isolation.

Let $\kappa(T, G, s) \in \{0, 1\}$ denote the separate *coverage condition*: G can consume $s_C(C)$, emit and decode $s_Y(Y)$, and its pretrained target prior has support for valid structures in that serialized target domain. Interface compatibility and target knowledge are both required: placing a field in an independently trained latent space may preserve it perfectly while still giving the frozen generator no meaning for its coordinates. When $\pi\kappa = 1$, the original task has the

generative reduction

$$\hat{z}_Y \sim p_\theta(z_Y \mid s_C(C), t), \qquad \hat{Y} = d_Y(\hat{z}_Y), \tag{19}$$

where t is optional text and d_Y is the declared target decoder. Solving this serialized conditional-generation problem solves T up to $\sim_T$. This is the formal content of the Unified Law; generation, estimation, control, reconstruction, and cross-property translation differ in the fields placed on the two sides of Eq. 19.

Backbone scope and learning prediction. The known fields C, requested fields Y, and directed relation $T : C \to Y$ describe the task's basic behavior. They do not change when a model uses autoregression, diffusion, flow matching, or another prediction mechanism. A backbone G enters through target coverage and the measured quality and acquisition cost. For another backbone that can consume the condition and generate the target representation, the same comparison applies, although its absolute quality, cost, and best route may change.

Let $Q_T(B; G, s)$ be native task quality under acquisition budget B, and let $\mathcal{C}_T(q; G, s)$ be the cost of reaching comparison quality q. If s^+ preserves more task-relevant information and better matches target knowledge already present in G than s^-, the law predicts

$$\begin{aligned} Q_T(B; G, s^+) &\gtrsim Q_T(B; G, s^-), \\ \mathcal{C}_T(q; G, s^+) &\lesssim \mathcal{C}_T(q; G, s^-). \end{aligned} \tag{20}$$

At fixed budget the aligned design should learn more; at matched quality it should cost less. The comparison keeps the task, generator, data, and evaluator fixed, and each task retains its native quality measure. The size of the advantage varies with target abstraction, information missing from the condition, and target knowledge already present in the backbone. The modality experiments and their plain-language interpretation appear in Sec. 14.10.

Three complementary levels of evidence. Section 5.3 supports the law at three levels. The first asks whether a rule formed from four initial Tunnels still works when applied at scale; it yields 1,148 trained and evaluated tasks. The second asks whether these routes follow each individual input; deliberately pairing their fixed predictions with the wrong targets increases error in every interface. The third asks why representation choice changes learning difficulty. We write the same surface-normal target in five lossless coordinate systems. All five reconstruct exactly the same image, while only the native VAE coordinates use the organization learned by the frozen generator during pretraining. Native coordinates reduce checkpoint-averaged angular error by 20.1%–26.6%, reach the lowest final error, and win 59 of 64 final sample–transformation comparisons.

The result gives Eq. 20 a direct experimental meaning: complete information is necessary, and a representation already understood by the backbone makes that information substantially easier to learn. The law therefore provides both a unified description of visual tasks and a practical rule for selecting their representations. Native visual coordinates are the first choice for dense visual targets; sparse and abstract targets call for a visual reorganization or a suitable output boundary.

The WSM realization additionally installs each route independently around frozen shared parameters. Dense Channel Heads use a 12.5M-parameter target-side tunnel in the reported configuration, while sparse and categorical targets add their declared native output components. Complete inference through the native target interface measures Q_T; a private codec round trip or partial preview does not establish output quality.

Formal consequences of the experiments. The preservation and coverage conditions above summarize a larger parameterized knowledge system derived in Sec. 14.11. Four consequences are especially important for interpreting the law.

First, representational accessibility and operational mastery are different quantities. For field X_i, define accessibility at tolerance ϵ_i as

$$A_i(\epsilon_i) = \mathbf{1}[r_i \le \epsilon_i], \tag{21}$$

and summarize operational mastery over a declared task set $\mathcal{T}_i$ by the task-quality profile

$$\mathbf{M}_i(\mathcal{T}_i) = (Q_T)_{T \in \mathcal{T}_i}, \tag{22}$$

where X_i participates as a condition, target, or intermediate field in one-to-one, one-to-many, many-to-one, or composed conversions. Accessibility is the first requirement; operational mastery adds measured relations with other fields. An accurate round trip does not imply high Q_T. The full experimental interpretation is given in Sec. 14.10, especially the distinction stated at page 211.

This distinction includes the earlier observation that reconstruction and compatibility are different quantities. The private round-trip error r_i of a codec can approach zero while the directed cross-field error $e_{i \to j}$ remains high; equal dimensions and matched marginal statistics do not make two latent coordinates synonymous (Eq. 26). Compatibility requires the preservation condition plus a sample-level relation and usable target semantics. The abstraction discarded by a codec is measured by $\alpha(Z; X) = H(X \mid Z)$, while the unresolved information in a directed task is $\mu(C \to Y \mid t) = H(Y \mid C, t)$. These quantities explain why the useful objective changes from continuous reconstruction toward symbolic prediction as abstraction increases, and why reversing condition and target creates a different task. Text, noise, another field, or a pretrained prior must supply choices not determined by the condition (Eqs. 27–29).

Second, representation design determines which distinctions survive. For a sparse field with occupied fraction

$\rho(Y)$, an empty image can have pixel loss $\mathcal{O}(\rho(Y))$ while missing every task-relevant object. Dense image codecs and sparse structured boundaries are therefore not interchangeable (Eq. 30). After information survives the boundary, path geometry and target knowledge remain separate resources: a shorter or straighter path does not imply a better decoded target when the frozen generator lacks support for valid target structures (Eq. 31).

Third, scalability is defined through marginal installation and terminal behavior. The interference quantity $\Delta Q_{i \leftarrow j}$ measures whether adding route j changes installed task i; independent installation requires it to remain zero within evaluation tolerance. Resident parameters grow with the number of trained routes. Each request activates task-specific parameters from only the selected route (Eqs. 32–33). Fourth, decreasing a surrogate training loss does not imply that the complete decoded output improves in its native task space (Eq. 34). These distinctions turn the Unified Law from a verbal analogy into a sequence of measurable claims about information, direction, prior knowledge, interference, cost, and final output.

Experimental origin of the conditions. The experiments above led to two law conditions and one installation criterion. Preservation arose from sparse-pixel, semantic-color, shared-VAE, and object-structure failures; Sec. 14.9 gives the information argument, and Sec. 14.10 explains how abstraction and direction change the useful learning form. Coverage arose when private latent spaces reconstructed themselves yet remained unreadable to a frozen generator, and when optimal transport, reflow, and stronger bridges changed the path without supplying target knowledge. Independent installation arose when joint lanes, fused conditions, packed LoRAs, and shared-target LoRAs coupled otherwise distinct capabilities. The complete chronological account is given in Sec. 14.8; Tables 18 and 19 record the corresponding families and representative variants. Table 20 gives the family-level map, while Table 21 lists the 198 currently indexed named variants separately with their direct observations and formal links. Section 14.13 summarizes the full record as eight core experimental principles.

These boundaries were recovered across 60 experimental families, 198 named variants, and at least 2,293 documented training attempts, then applied in WSM v1. All 1,148 reported tasks were trained and evaluated through the same task-keyed acquisition framework, with interface-specific components declared explicitly. The matched MTA–NJA experiments are mechanism-level evidence within this realization—19.0–38.5$\times$ faster structured acquisition under MTA and an 11.0% lower RGB error under NJA—not part of the law's definition. Likewise, the independent-annotation evaluation tests whether acquired routes transfer beyond the annotation convention without retraining.

Refutation criteria. The representational statement would fail if a task satisfying the declared preservation and coverage conditions could not be expressed and evaluated through Eq. 19. The learning prediction would fail if a matched comparison reliably reversed the ordering in Eq. 20. Independent installation would fail if adding a route degraded an existing route. Its claimed breadth would fail if such counterexamples accumulated throughout the stated visual regime. A failure at a disclosed boundary—erased information, an unreadable target space, an uncovered target prior, or absent paired supervision identifies which premise is missing; it does not silently redefine the task as a successful case. These criteria make the law, its empirical realization, and the limits of WSM v1 separately testable.

Computer vision is ready for an organizing law of its own. After decades of progress measured mainly as separate models and benchmarks, the field needs a principle that explains how those tasks relate and predicts how new ones can be constructed. Scaling laws describe how performance changes as resources grow; the Unified Law addresses what can be scaled—the modular visual capability space itself. The four interfaces make a concrete and refutable prediction: once task-relevant information is encoded as World Snapshots, heterogeneous visual computations and newly aligned visual information can be incorporated into the same generative formulation. WSM v1 provides the first thousand-task realization and unusually broad evidence for this principle. More broadly, WSM and LLMs can serve as peer foundations. An LLM unifies language and abstract reasoning; WSM unifies perception, generation, reconstruction, and control over explicit visual world states. Coupling these foundations could lead toward a general-purpose multimodal foundation model for universal task execution, in which linguistic intelligence and visual world intelligence operate as coequal, mutually conditioning capabilities. By these criteria, the Unified Law is a testable law-scale hypothesis supported by the first thousand-task realization. It opens a path from task-specific models toward general-purpose visual world intelligence.

8. Conclusion

We presented the Unified Law for Visual Tasks, which expresses visual tasks as transformations among image-like World Snapshots. World Snapshot Graphics encodes RGB and auxiliary attributes in a representation with up to 256 channels, while four interfaces express generation, estimation, conditional control, and reconstruction in one generative formulation. Together, these interfaces and TunnelingNet realize 1,148 evaluated visual tasks within a dense graph of 3,042 configured mappings. We validate both the method and the Unified Law at this scale, establishing their effectiveness across more than one thousand tasks.

Our experiments show that TunnelingNet preserves the pretrained RGB-synthesis prior while efficiently acquiring structured visual transformations, reaching up to approximately $100\times$ greater training efficiency than the parameter-matched native-attention baseline. Across all evaluated tasks, WSM also achieves consistently strong performance, surpassing specialized prior methods—including Best Paper and Best Paper Candidate systems—on the reported metrics. The mathematical scalability of the Unified Law further connects model growth to task growth: aligned modalities expand direct-task breadth quadratically, visual chains expand compositional depth exponentially, and independent task routes can be trained in parallel. These routes are registered parameters of one modular WSM and are sparsely activated by an explicit task key; independent optimization and physical sharding do not make them standalone models. WSM therefore provides one extensible generative foundation for a large and continually growing vocabulary of visual tasks. Its longer-term trajectory is to unify, expand, and ultimately transcend the fixed task boundaries of conventional Computer Vision.

In addition, we compare WSM with the visual tasks established before this work and obtain two findings. First, among the tasks recorded by Papers with Code, 1,477 are visual or vision-related tasks, broadly representing the vast established field of Computer Vision. All 1,477 tasks are compatible with the WSM interfaces and can be expressed by the same system. **After half a century of development, we bring the vast field of Computer Vision into a unified generative framework.** Second, **WSM creates 655–1,056 entirely new visual tasks and expands Computer Vision by 44.3–71.5%.** Every evaluated task is accompanied by a trained baseline. Together, these results establish WSM as a task-generative foundation that consolidates the historical field and opens a substantially larger visual-task space.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165. 2

[2] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. 2

[3] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 2

[4] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 2

[5] Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation, 2025. URL https://arxiv.org/abs/2502.20388. 2

[6] Xinlong Wang et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. URL https://arxiv.org/abs/2409.18869.

[7] Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Bridging continuous and discrete tokens for autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2

[8] Juan Maldacena. The large-n limit of superconformal field theories and supergravity. *International Journal of Theoretical Physics*, 38(4):1113–1133, 1999. ISSN 1572-9575. doi: 10.1023/A:1026654312961. URL https://doi.org/10.1023/A:1026654312961. 2, 6

[9] Chris Fields, James F. Glazebrook, and Antonino Marcianò. The physical meaning of the holographic principle. *Quanta*, 11:72–96, November 2022. ISSN 1314-7374. doi: 10.12743/

quanta.v11i1.206. URL http://dx.doi.org/10.12743/quanta.v11i1.206.

[10] M. Hossein Eybposh, Nicholas W. Caira, Mathew Atisa, Praneeth Chakravarthula, and Nicolas C. Pégard. Deepcgh: 3d computer-generated holography using deep learning. *Opt. Express*, 28(18):26636–26650, Aug 2020. doi: 10.1364/OE.399624. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-28-18-26636. 2, 6

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 27, 30, 34, 172

[12] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. 2, 172, 190

[13] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538. 2, 172

[14] Danyang Zhang, Junhao Song, Ziqian Bi, Xinyuan Song, Yingfang Yuan, Tianyang Wang, Joe Yeong, and Junfeng Hao. Mixture of experts in large language models, 2025. URL https://arxiv.org/abs/2507.11181. 172

[15] Mingqiao Ye, Zhaochong An, Zhitong Gao, Xian Liu, François Fleuret, Chuan Li, Amir Zadeh, Serge Belongie, Afshin Dehghan, Jesse Allardice, David Mizrahi, Oğuzhan Fatih Kar, Roman Bachmann, and Amir Zamir. MODUS: Decoder-only any-to-any modeling of diverse modalities. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2607.25948. 27, 30, 34

[16] Rahul Ravishankar, Zeeshan Patel, Jathushan Rajasegaran, and Jitendra Malik. Scaling properties of diffusion models for perceptual tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 12945–12954, June 2025. 2, 27, 30, 34

[17] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 27, 30, 34, 172

[18] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 2, 172

[19] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4646.

[20] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. Lorashop: Training-free multi-concept image generation and editing with rectified flow transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5596.

[21] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17281–17291, October 2025. doi: 10.1109/ICCV51701.2025.01605. 2, 172

[22] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator, 2024. URL https://arxiv.org/abs/2405.07510. 6, 189

[23] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[24] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

[25] Pengsheng Guo and Alexander G. Schwing. Variational rectified flow matching, 2025. URL https://arxiv.org/abs/2502.09616.

[26] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742. 6, 190

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003. 6, 188

[28] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation, 2024. URL https://arxiv.org/abs/2309.06380.

[29] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747. 6

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 6, 210

[31] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025. 6, 9, 190

[32] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang,

Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6, 190

[33] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103. 6, 190

[34] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2025. 6

[35] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328. 6

[36] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 6, 191

[37] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4256–4265, 2026. 6

[38] Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners, 2026. URL https://arxiv.org/abs/2604.20329. 6

[39] Houyuan Chen, Hong Li, Xianghao Kong, Tianrui Zhu, Shaocong Xu, Weiqing Xiao, Yuwei Guo, Chongjie Ye, Lvmin Zhang, Hao Zhao, and Anyi Rao. Unividx: A unified multimodal framework for versatile video generation via diffusion priors. *ACM Transactions on Graphics*, 45 (4), July 2026. doi: 10.1145/3811304. URL https://doi.org/10.1145/3811304. 6, 190

[40] Ruohao Guo, Xianghua Ying, Yaru Chen, Dantong Niu, Guangyao Li, Liao Qu, Yanyu Qi, Jinxing Zhou, Bowei Xing, Wenzhen Yue, Ji Shi, Qixun Wang, Peiliang Zhang, and Buwen Liang. Audio-visual instance segmentation. *arXiv preprint arXiv:2310.18709*, 2023. 7

[41] Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, and Guosheng Lin. Magicarticulate: Make your 3d models articulation-ready, 2025. URL https://arxiv.org/abs/2502.12135. 7

[42] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. URL https://arxiv.org/abs/1512.03012. 7

[43] Giuseppe Vecchio and Valentin Deschaintre. Matsynth: A modern pbr materials dataset. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 22109–22118. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.02087. URL http://dx.doi.org/10.1109/CVPR52733.2024.02087. 7

[44] Denis Zavadski, Damjan Kalšan, and Carsten Rother. Primedepth: Efficient monocular depth estimation with a stable diffusion preimage. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 922–940, December 2024. 7

[45] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02073. 7

[46] Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. Normalcrafter: Learning temporally consistent normals from video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, October 2025. 7

[47] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851. 7, 37

[48] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks, 2016. URL https://arxiv.org/abs/1612.01925. 7

[49] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 7

[50] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. URL https://arxiv.org/abs/1906.08172.

[51] Baldomero R. Árbol and Dan Casas. Bodyshapegpt: Smpl body shape manipulation with llms, 2024. URL https://arxiv.org/abs/2410.03556. 7

[52] Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, and Guosheng Lin. Magicarticulate: Make your 3d models articulation-ready. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 15998–16007, June 2025. 7

[53] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 7

[54] Manolis Savva, Angel X. Chang, and Pat Hanrahan. Semantically-Enriched 3D Models for Common-sense Knowledge. *CVPR 2015 Workshop on Functionality, Physics, Intentionality and Causality*, 2015. 7

[55] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, June 2024. 27, 30, 34, 37

[56] Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7861–7871, June 2024. 27, 30, 34

[57] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 241–258, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72670-5. doi: 10.1007/978-3-031-72670-5_14. 27, 30, 34

[58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 5294–5306, June 2025. 27, 30, 34

[59] Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, Jinsheng Quan, Zhongang Cai, Haiwen Diao, Ziwei Liu, Lei Yang, Dahua Lin, and Quan Wang. Vision as unified multimodal generation, 2026. URL https://arxiv.org/abs/2607.06560. 27, 30, 34, 177

[60] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226. 27, 30, 34

[61] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42961–42992. Curran Associates, Inc., 2023. doi: 10.52202/075280-1862. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/862f45ccecb2275851bc8acebb8b4d65-Paper-Conference.pdf. 27, 30, 34

[62] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0491. 27, 30, 34

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cFTi3gLJ1X. 37

[64] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 37

[65] World Snapshot. Papers with Code Data Archive. https://github.com/World-Snapshot/papers-with-code, 2025. Final archived Papers with Code snapshot preserved before the service ended on 24 July 2025; commit baee722f75e710c5eaec44bb6f3f9630c212d636. 71, 103

[66] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361. 72

[67] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Trans. Graph.*, 43(4), July 2024. ISSN 0730-0301. doi: 10.1145/3658150. URL https://doi.org/10.1145/3658150. 84, 186, 225

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=u1cQYxRI1H. 84, 225

[69] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114. 186

[70] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 186

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762. 188

[72] Ollin Boer Bohan. Taesd: Tiny autoencoder for stable diffusion. https://github.com/madebyollin/taesd, 2023. GitHub repository. 189

[73] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 189

[74] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239. 189

[75] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17191–17202. IEEE, 2025. 190

[76] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 192

[77] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 906–915. IEEE, 2024. 193

[78] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 193

[79] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 193

[80] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720, 2023. 193

[81] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *European conference on computer vision*, pages 398–416. Springer, 2022. 194

[82] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629. IEEE, 2023. 197

[83] Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. 197

[84] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 199

[85] aigc apps. Videox-fun: A video generation pipeline for diffusion transformer, 2026. URL https://github.com/aigc-apps/VideoX-Fun. 199

[86] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 201

[87] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pages 35791–35820, 2026. 226

UNIFIED LAW FOR VISUAL TASKS (THE FOUR LAWS OF COMPUTER VISION)

Supplementary Material

9. Appendix: Implementation Details

Task definitions. Two JSON files define the dense and sparse tasks. Each task specifies its Channel Head index, task type, input modalities, target modality, implementation type, and the availability of paired training data. The dense file configures 3,042 visual task mappings around one shared generative backbone, while the sparse graph contains 11 additional digitized mappings. Sparse Heads are reported separately in this appendix and are not included in the main-paper count of 1,148 tasks. The frozen benchmark provides checkpoints and complete results on the evaluation set for 1,148 dense Channel Heads. The remaining mappings use the same input–output format and define the next set of tasks to train and evaluate. The training program reads these definitions, prioritizes selected task types, and trains each Channel Head on one GPU.

Dense WSG path. Dense channels are trained from latent caches. For each head, the launcher passes the target modality, condition modality list, Channel Head index, and task type into the TunnelingNet trainer. The dataset intersects target and condition sample IDs, applies optional dataset-family whitelists, and builds aligned tuples of target latent, condition latent, reference latent, caption embedding, frame index, video id, and source metadata. Reference routing is task dependent. The selected text-guided narrow-generation recipe retains NJA and routes first-frame RGB through the native appearance-reference slot. Inverse rendering routes RGB as an extra condition. Audio-conditioned dense masks append projected AST audio features as an additional condition lane.

Task recipes. The implementation resolves a recipe from the task type. Estimation/reconstruction, conditional control, and inverse rendering use Modality Tunneling Attention with a trainable target lane and frozen condition lanes. Text-guided narrow generation uses Native Joint Attention; in the selected recipe, first-frame RGB enters the native appearance-reference slot. Each experiment saves its final attention mode, trainable lanes, branches, normalization rules, optimization and validation settings, and Channel Head metadata in `resolved_runtime_config.json`.

Optimization and monitoring. Dense heads are optimized with AdamW on the trainable tunnel parameters only. Training uses bfloat16, gradient checkpointing when requested, gradient clipping, fixed validation timestep grids, validation-loss monitoring, and early stopping based on validation improvement. Preview generation decodes the predicted target, the condition, and the ground truth through the VAE. For WSG modalities, the preview path writes decoded values back to a 256-channel WSG array and invokes the WSG reader so that depth, normal, pose, mask, material, and audio-related targets are inspected in their native visualization format.

Table 11. **Final categorical output metrics by target.** Strict accuracy, dominant-class accuracy (DAcc), and mIoU are measured on all evaluated categorical Channel Heads.

Target	Name	Heads	Samples	Strict	DAcc	mIoU
111	Object category	39	497	0.250	0.288	0.255
131	Material class	39	460	0.311	0.298	0.298
135	Container flag	39	429	0.915	0.925	0.493
136	Object class	39	460	0.141	0.085	0.085
195	Semantic segmentation	22	528	0.341	0.401	0.107

Normalization and reversibility. Each WSG config carries a portable set of normalization rules with channel ranges and per-modality modes. Training uses normalized tensors, while visualization and export invert these rules before invoking WSG readers. This keeps the learning problem numerically stable and preserves the original physical or semantic units at the interface boundary.

Sparse WSG path. Sparse Heads follow the same input–output task definition but use task-specific adapters before returning to the WSG convention. The sparse launcher resolves entries from the sparse JSON graph and dispatches by `adapter_task`. Human structure heads build RGB-crop datasets and cache PARE features for SMPL/SMPLer-X, DWPose, and MediaPipe targets. The bounding-box head uses full RGB images with visibility and area filters. The object-structure head converts WSG object geometry into normalized point clouds and predicts a breadth-first skeleton tree with parent slots and relative or absolute coordinates. Audio heads build frame-aligned audio features from WSG audio channels; the visual-to-audio branch can adapt MusicGen with image-prefix conditioning through the same compact-tunnel principle. The sparse system therefore handles lists, boxes, keypoints, skeletons, body parameters, and audio tokens as digitized targets in the same directed WSG graph.

Parallel training. Dense Channel Heads can be trained independently: the scheduler assigns them to available GPUs, uses one process per Head, saves each experiment in `train_log/Channel_Head_xxxxxx`, and resumes or skips completed Heads. Sparse training uses the same Channel Head indexing scheme. Each adapter owns its data builder, cache, and trainer. Dense WSG targets exploit the shared rectified-flow generator; sparse digitized targets use compact supervised adapters when their native output space is structured rather than image-like.

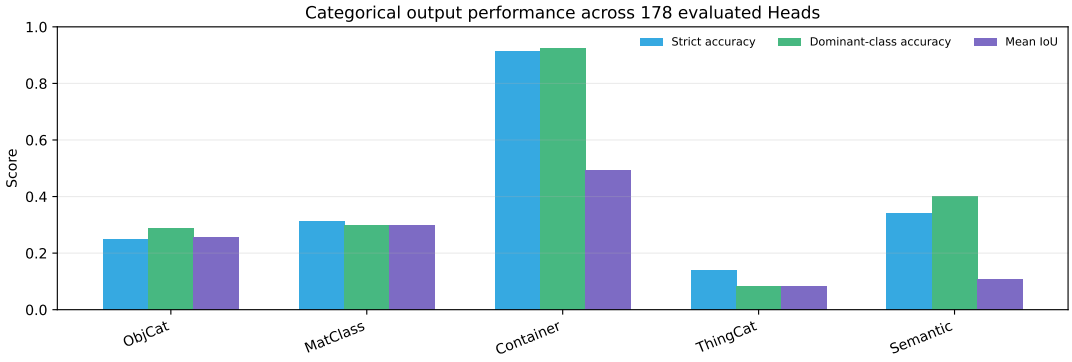


Figure 48. **Categorical output performance across 178 complete Channel Heads.** Bars report the final composed output for strict pixel accuracy, dominant-class accuracy, and mean IoU across all five semantic categorical targets.

Table 12. **Evaluation of the seven default native sparse Channel Heads.** Each Head is evaluated in its native output space, so its metrics follow the predicted structure. All values use the selected model for that Head.

Head	Mapping	Native metrics	Result
000000	RGB crop → SMPL pose and shape	SMPL shape-coefficient MAE ↓	0.0373
000001	RGB → person boxes	Precision / Recall / F1 ↑	0.820 / 0.770 / 0.794
000002	RGB crop → DWPose	Median XY error ↓ / PCK@0.10 ↑ / PCK@0.20 ↑	0.0925 / 0.526 / 0.725
000003	RGB crop → MediaPipe pose	Median XY error ↓ / PCK@0.10 ↑ / PCK@0.20 ↑	0.0852 / 0.568 / 0.797
000005	Audio + RGB → source box	Mean IoU / Macro-category IoU / IoU@0.5 ↑	0.325 / 0.325 / 0.351
000006	Audio → semantic category	Top-1 / Balanced accuracy ↑	0.671 / 0.602
000010	RGB + Audio → sounding regions	Region-selection F1 / Semantic-category accuracy ↑	0.768 / 0.796

10. Appendix: Extended Experimental Analysis

10.1. Ablations and Categorical Results

Categorical output protocol. Categorical targets expose a stricter requirement than continuous channels: the generated support must be spatially correct and every supported region must also receive the correct semantic identity. The classification branch is therefore an integral component of each categorical Channel Head. We evaluate only the final composed prediction—generated spatial support followed by category assignment—rather than treating the intermediate pseudo-color field as a separate model variant. Across five categorical targets, this evaluation covers 178 complete Heads and 2,374 evaluation outputs.

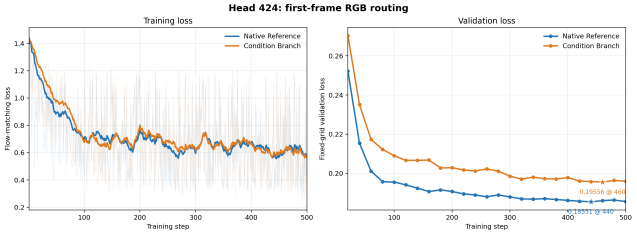


Figure 49. **First-frame RGB routing for Edge-to-RGB generation.** Head 424 routes the same first-frame RGB through either the native FLUX reference branch or the WSM first-frame condition branch. The two routes follow similar optimization trajectories. At update 500, native-reference routing obtains a validation loss of 0.1856, compared with 0.1959 for the condition branch, while their smoothed training losses are 0.5838 and 0.5633, respectively. The recommended Head 424 therefore uses the native reference route as its final training setting.

Additional RGB-routing ablation. Figure 49 complements the main-paper routing study with Head 424, which generates RGB from text, Edge, and first-frame RGB. We isolate how the first-frame RGB enters the model while holding the 500-update training budget fixed.

Aggregate categorical performance. Figure 48 and Table 11 report the complete categorical results. The binary container target is already reliable, reaching 0.915 strict accuracy and 0.925 dominant-class accuracy. Object category and material class obtain mIoU values of 0.255 and 0.298. Fine-grained object-class and semantic-segmentation outputs remain harder, with mIoU values of 0.085 and 0.107. Strict accuracy, dominant-class accuracy, and mIoU are reported together because they measure overall pixel agreement, the identity of the principal semantic region, and both

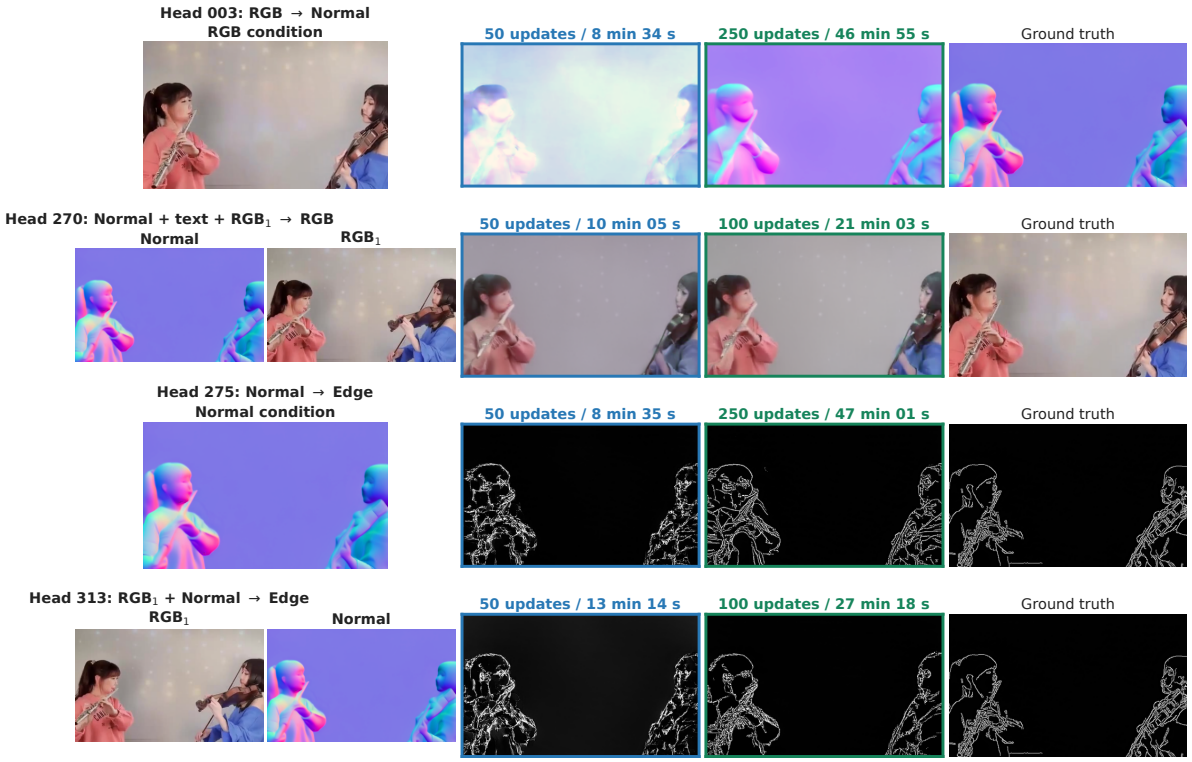


Figure 50. **Extended minute-scale acquisition evidence.** Each row shows the condition, the first saved MTA output at update 50, a later selected output, and the ground truth. With batch size one, update 50 corresponds to only 50 sampled examples and is reached in 8 minutes 34 seconds, 10 minutes 5 seconds, 8 minutes 35 seconds, and 13 minutes 14 seconds for Heads 003, 270, 275, and 313. The selected 250-, 100-, 250-, and 100-update outputs require 46 minutes 55 seconds, 21 minutes 3 seconds, 47 minutes 1 second, and 27 minutes 18 seconds, respectively. Blue and green borders mark the first saved and selected outputs.

missed and spurious semantic support, respectively.

Errors introduced by categorical representations. The qualitative artifact analysis applies an additional, representation-level filter. We omit MatRegion from qualitative panels because its colors encode per-object local region IDs without cross-sample semantic identity.

Categorical accuracy across input modalities. End-to-end categorical reliability is not yet uniformly strong and stable across 114 cross-modal mappings. This set is not a list of failed models: several mappings already provide useful or strong predictions, but their reliability remains sensitive to the semantic evidence available in the input. It contains 57 conditional-control and 57 RGB-assisted inverse-rendering Heads, with 38 mappings for each of Targets 111, 131, and 136.

Outputs that need further improvement. Targets 111, 131, and 136 require more robust separation of closely related categories across heterogeneous input modalities. Target 195 is retained in the reported categorical results after the region-handling correction, while Target 103 is evaluated as a dense human-segmentation map rather than as a categorical-refiner output. Future work will prioritize the three withheld target families through better-balanced classification, stronger region-level evidence, and systematic re-evaluation under the same evaluation protocol.

10.2. More Results

Minute-Scale Rapid Learning. The compact main-paper sequence isolates learning during the first 10–50 updates trajectory and its 200-update endpoint. Figure 50 retains the complementary evidence used during model development: the input condition, the first saved MTA output, a later selected model, and the ground truth. Because batch size is one, update count is also the number of sampled training examples. A Channel Head develops visible task structure after only tens of examples and reaches the selected quality within at most 250 examples and 47 minutes. Qualitative change was observed interactively at approximately five minutes, but no decoded checkpoint was written at that time; the figure therefore reports only saved outputs and measured wall-clock times.

Sparse-task results. The default sparse launcher selects seven verified Heads: SMPL pose and shape, person boxes, DWPose, MediaPipe pose, sound-source localization, audio category recognition, and audio-selected semantic regions. Table 12 evaluates each output using its native geometry or classification metric. Optional object and visual-to-audio experiments are not included in this default set. Since the sparse and custom graphs are compact, each validated Head stores the path of its selected model in its task entry; the 1,148 dense Heads use a separate model list.

10.3. Capability Showpiece: Speedrunning Celebrated Visual Systems

The preceding 1,148 Heads already establish the scalability of TunnelingNet. We therefore end not with another routine mapping, but with a deliberately ambitious showpiece: two signature capabilities that previously motivated celebrated standalone systems are added to the same frozen 4B backbone as ordinary lightweight Channel Heads. We target the transparent-asset capability of LayerDiffuse [67] and the illumination-controlled relighting task of the full-score ICLR 2025 IC-Light paper [68]. The transparency Head requires only 1,000 updates with batch size four (4,000 sampled training examples), and the relighting Head only 2,800 single-sample updates. Figures 51 and 52 show held-out 10-step inference: WSM completes a text-to-RGB-to-RGBA generation chain and performs spatially controlled relighting while preserving scene content. In TunnelingNet, headline capabilities that previously justified dedicated systems and papers become lightweight additions to one shared backbone. This is the practical power of deep tunneling: a strong new visual capability no longer requires building another model from scratch.

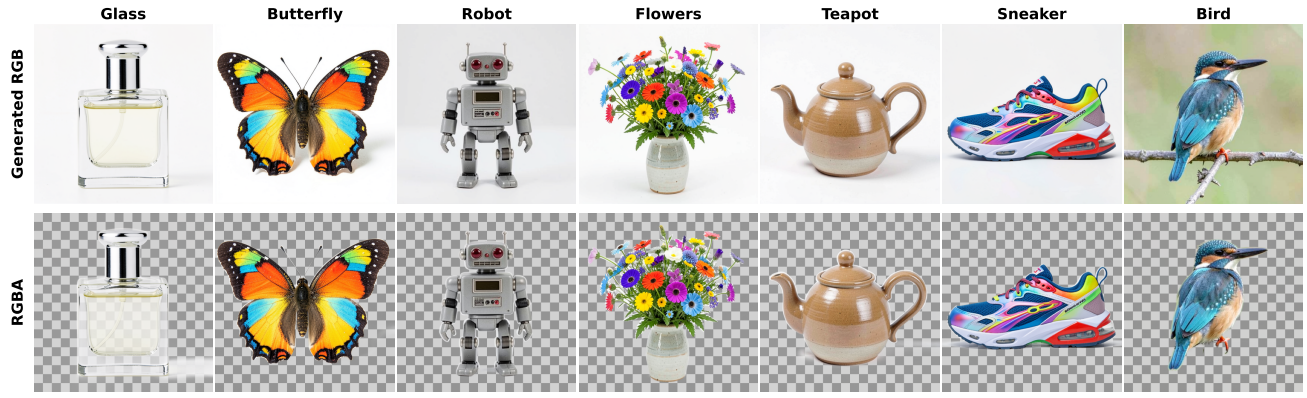


Figure 51. **Text-to-RGBA generation through a custom transparency Head.** Each column executes the complete chain rather than reading a cached transparent asset: WSM first generates the RGB image from text, predicts its alpha matte with the custom LayerDiffuse task Head, and composes the resulting RGBA image over a checkerboard. Fine structures such as antennae, flower stems, handles, laces, and feathers remain visible in the predicted transparency.

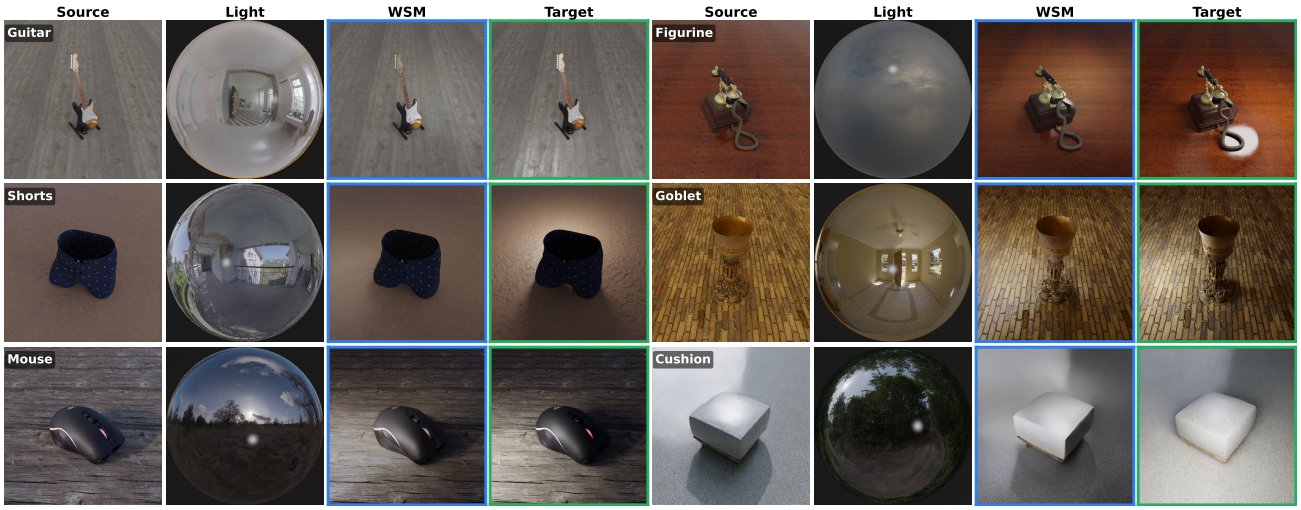


Figure 52. **Rapid acquisition of the full-score ICLR 2025 IC-Light relighting task.** Each group shows the source image, the spatial light probe, the WSM prediction, and the held-out target. Blue and green borders mark the WSM prediction and target, respectively. A single custom Head redirects the frozen backbone according to the requested illumination while retaining object identity, material, texture, and geometry.

11. Appendix: ControlWorld User Guide

ControlWorld is the reference interface for selecting and running WSM tasks. A user selects a trained mapping in the WSG Task Graph, supplies its image and text inputs, selects the corresponding trained model, and generates the output. The output panel then applies the target modality's WSG Reader visualization and legend. The desktop captures below show the same released 10-step Euler results used in the qualitative evaluation, with identical tasks, inputs, models, seeds, and outputs.

Reading each capture. The upper task graph identifies the selected input–output route. The input cards show the supplied conditions, the central controls expose checkpoint and sampling settings, and the output card shows the decoded result. Parts I, II, VII, and IX are reproduced here; for Part IX, only the first three source figures are included.

11.1. The Four WSM Interfaces

Part I exposes one desktop workflow for each of the four interface definitions.

11.2. RGB Estimation of Geometry and Motion

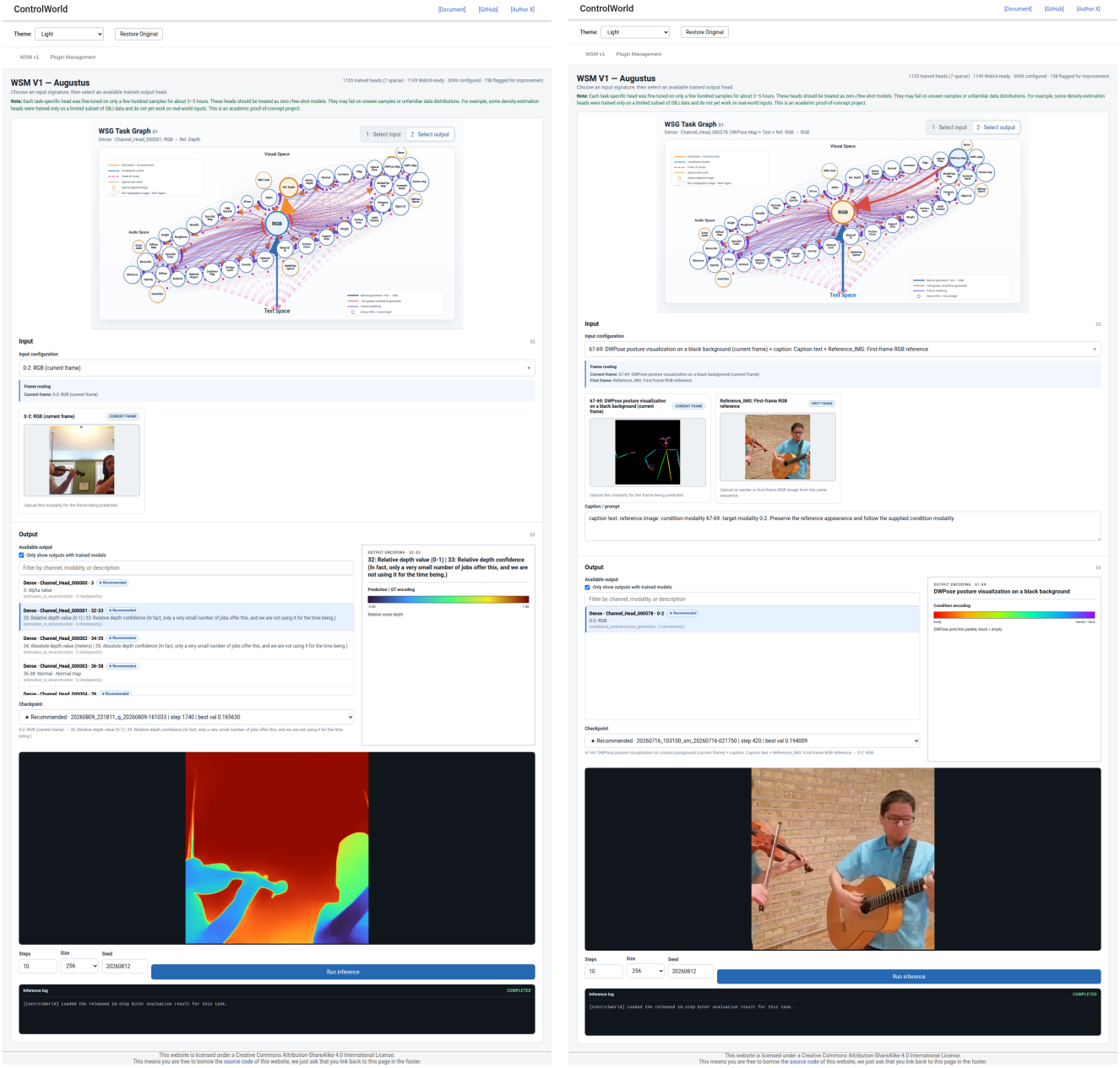
Part II demonstrates the complete RGB-centered geometry and motion sheet.

11.3. Narrow Generation from Geometry and Human Structure

Part VII uses the first four released narrow-generation examples, as a compact walkthrough of geometry-to-RGB operation.

11.4. Conditional Control

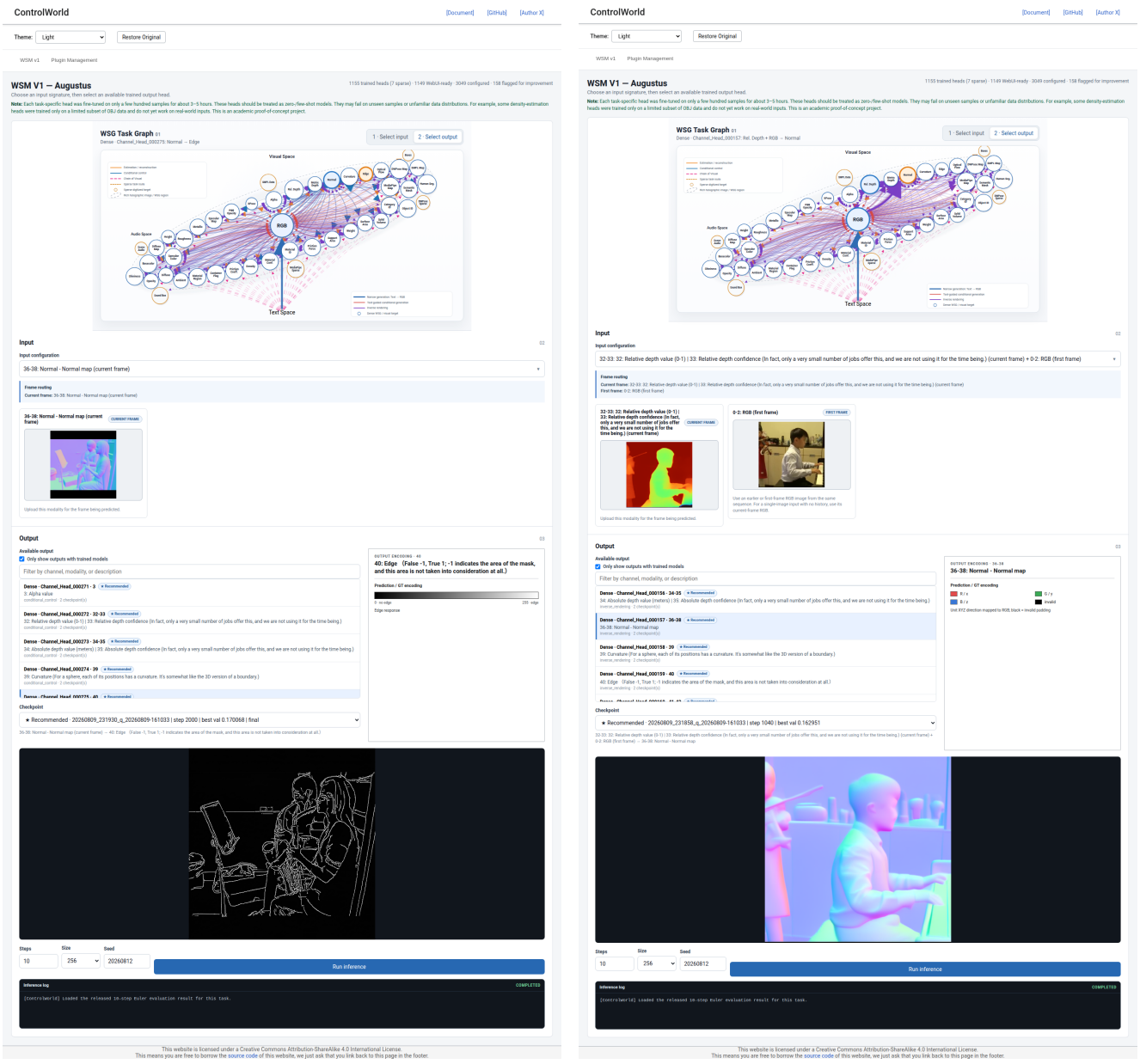
Part IX uses every example from its first three qualitative sheets (18 tasks) and omits sheets four through six.



(a) **Head 000001:** RGB → RelDepth.

(b) **Head 000578:** DWPose+Text+RGB1 → RGB.

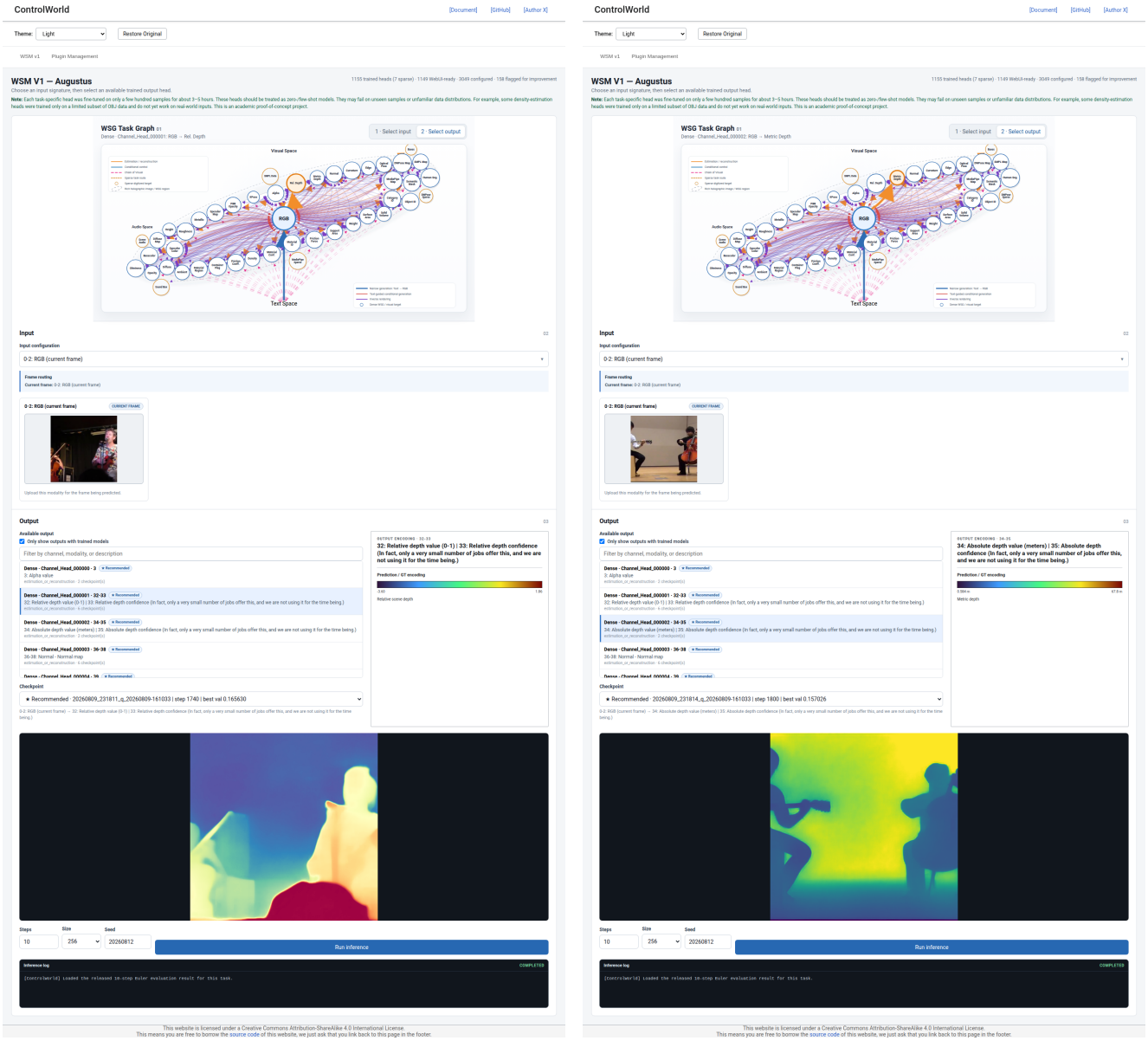
Figure 53. ControlWorld desktop workflows for Qualitative Part I: Head 000001 (RGB to RelDepth) and Head 000578 (DWPose+Text+RGB1 to RGB). Each panel preserves the full task graph, input, sampling, and decoded-output interface.



(a) **Head 000275:** Normal → Edge.

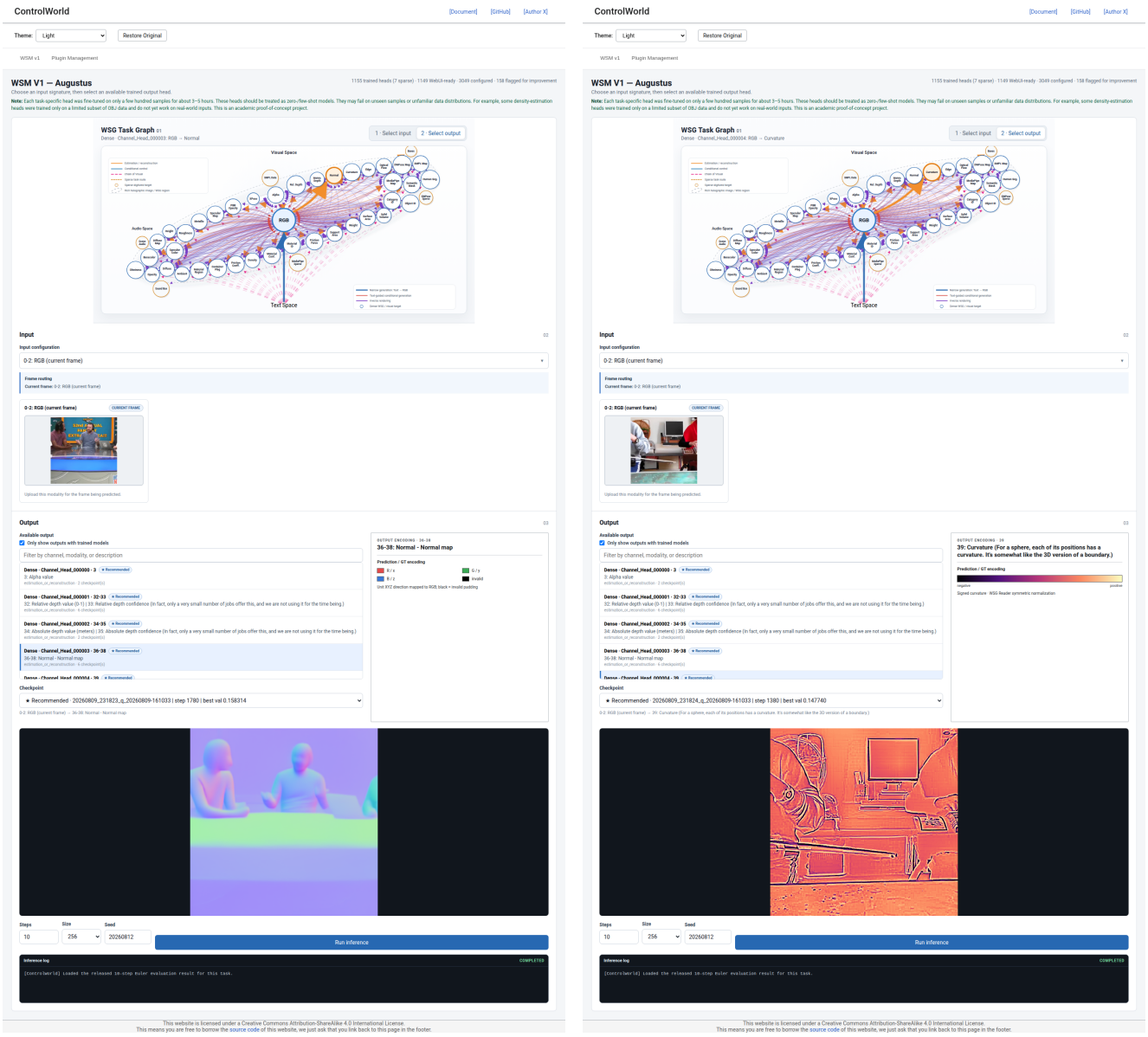
(b) **Head 000157:** RelDepth+RGB → Normal.

Figure 54. ControlWorld desktop workflows for Qualitative Part I: Head 000275 (Normal to Edge) and Head 000157 (RelDepth+RGB to Normal). Each panel preserves the full task graph, input, sampling, and decoded-output interface.



(a) **Head 000001:** RGB → RelDepth. (b) **Head 000002:** RGB → AbsDepth.

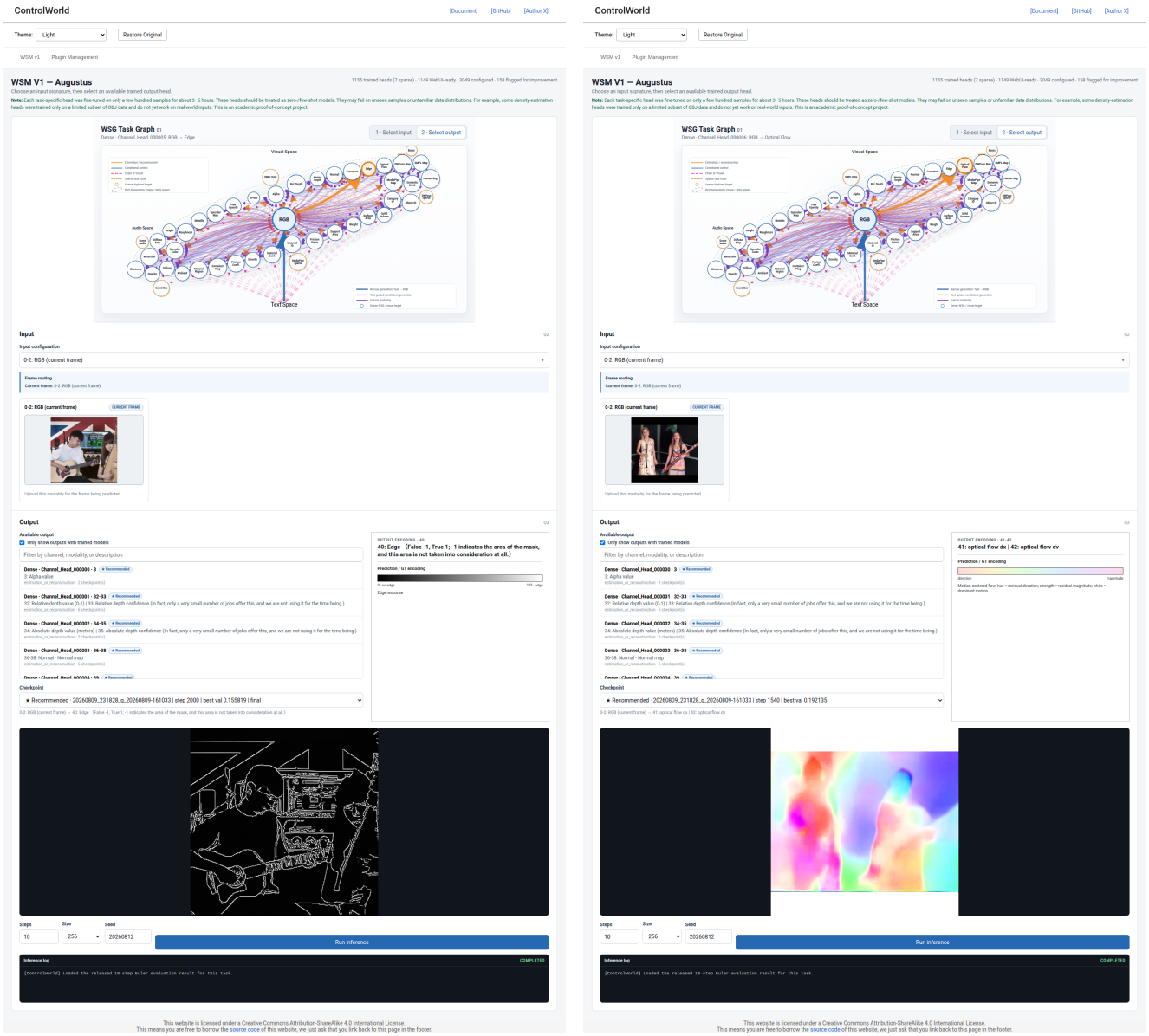
Figure 55. ControlWorld desktop workflows for Qualitative Part II: Head 000001 (RGB to RelDepth) and Head 000002 (RGB to AbsDepth). Each panel preserves the full task graph, input, sampling, and decoded-output interface.



(a) **Head 000003:** RGB → Normal.

(b) **Head 000004:** RGB → Curve.

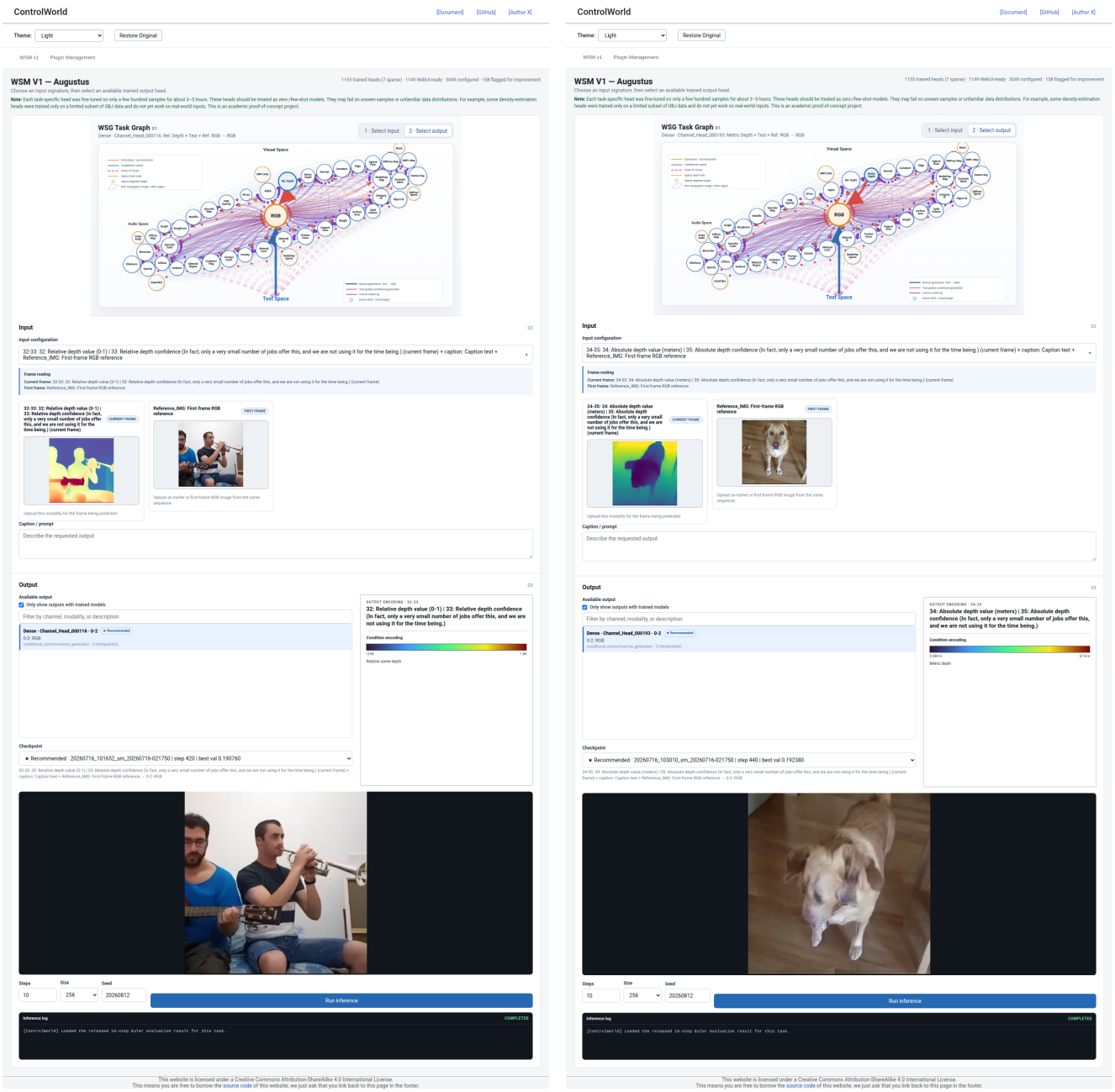
Figure 56. ControlWorld desktop workflows for Qualitative Part II: Head 000003 (RGB to Normal) and Head 000004 (RGB to Curve). Each panel preserves the full task graph, input, sampling, and decoded-output interface.



(a) **Head 000005:** RGB → Edge.

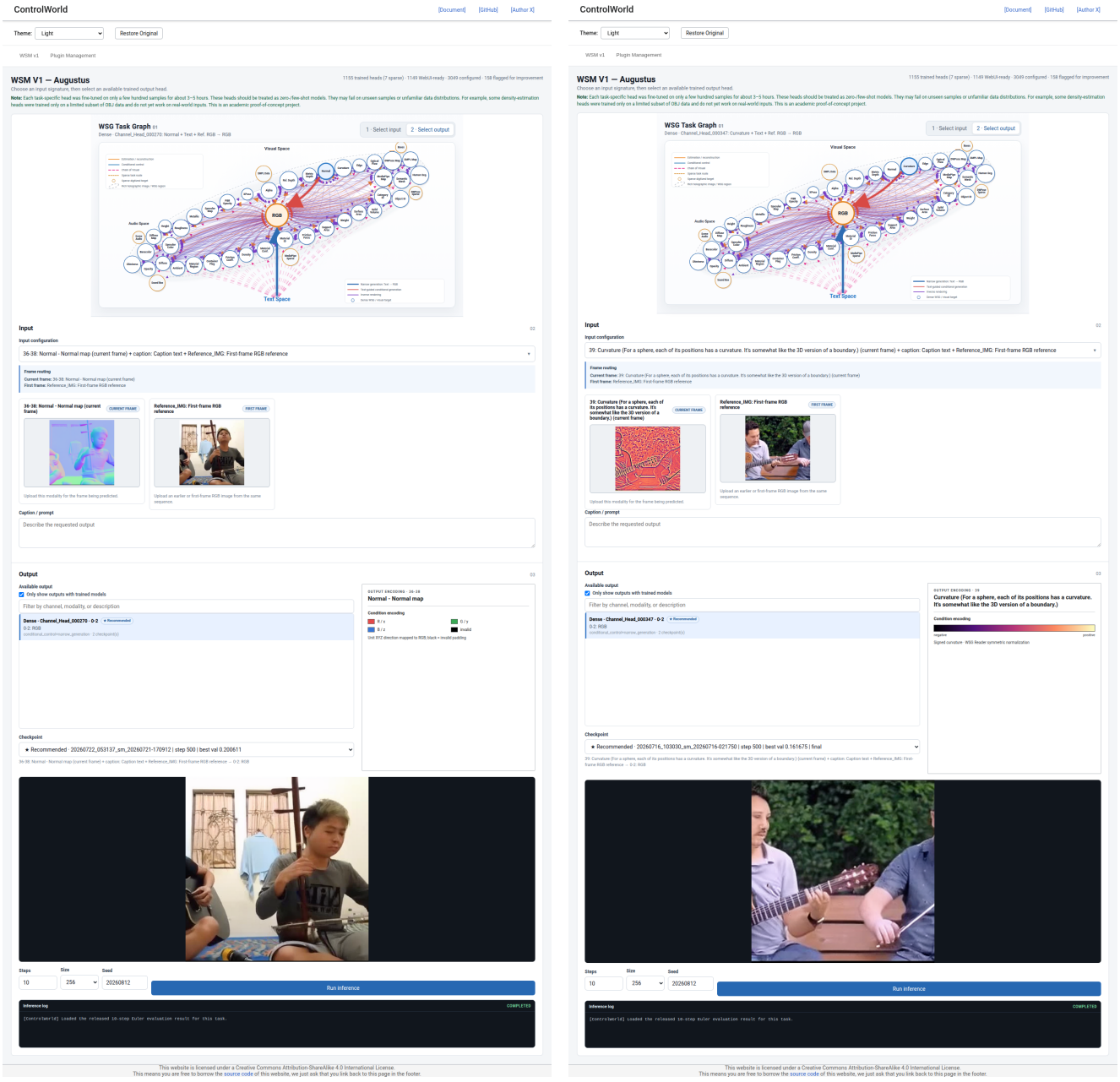
(b) **Head 000006:** RGB → Flow.

Figure 57. ControlWorld desktop workflows for Qualitative Part II: Head 000005 (RGB to Edge) and Head 000006 (RGB to Flow). Each panel preserves the full task graph, input, sampling, and decoded-output interface.



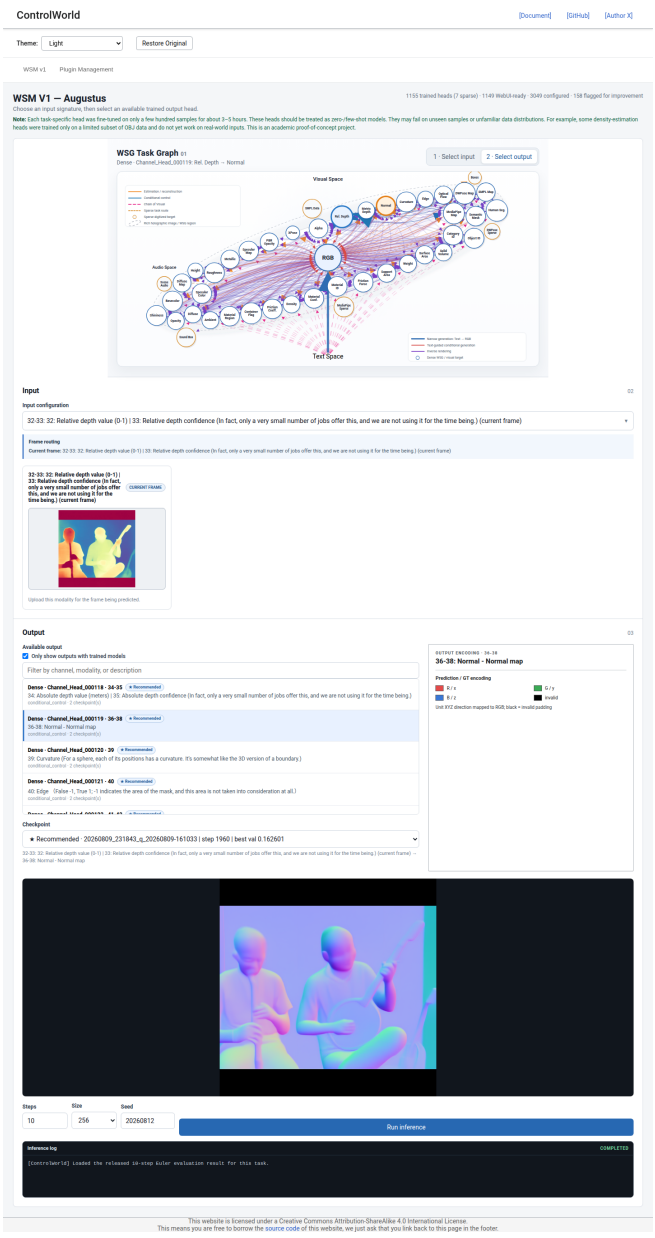
(a) **Head 000116:** RelDepth+Text+RGB1 → RGB. (b) **Head 000193:** AbsDepth+Text+RGB1 → RGB.

Figure 58. ControlWorld desktop workflows for Qualitative Part VII: Head 000116 (RelDepth+Text+RGB1 to RGB) and Head 000193 (AbsDepth+Text+RGB1 to RGB). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

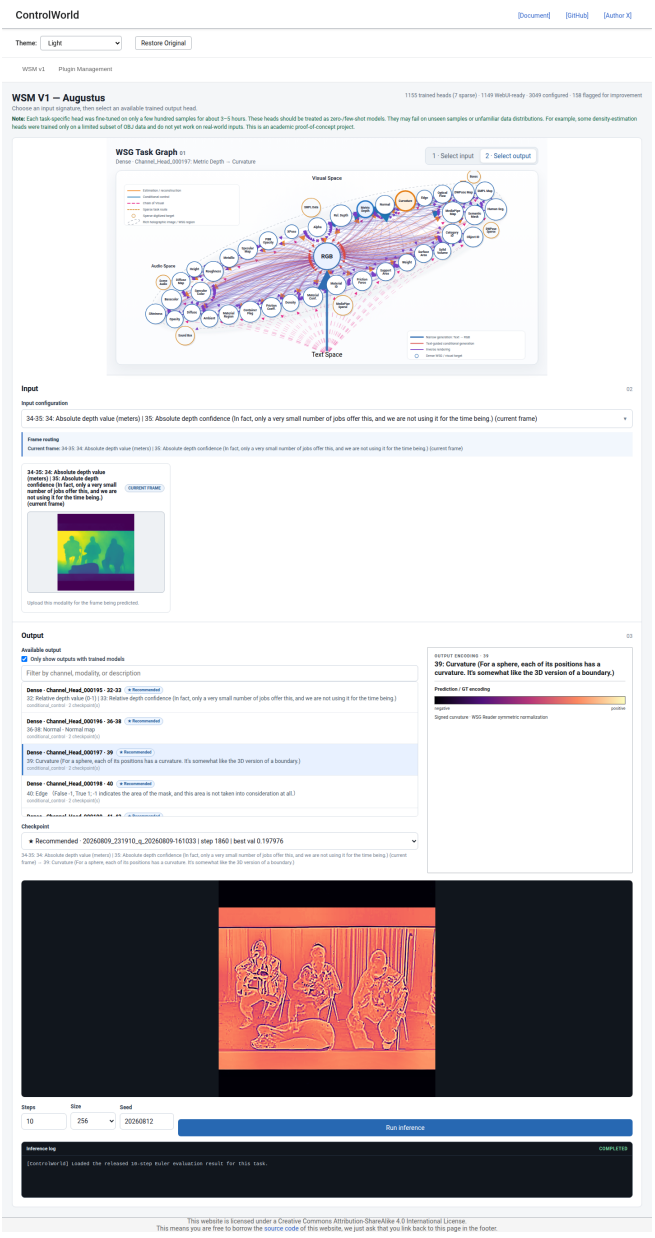


(a) **Head 000270:** Normal+Text+RGB1 → RGB. (b) **Head 000347:** Curve+Text+RGB1 → RGB.

Figure 59. ControlWorld desktop workflows for Qualitative Part VII: Head 000270 (Normal+Text+RGB1 to RGB) and Head 000347 (Curve+Text+RGB1 to RGB). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

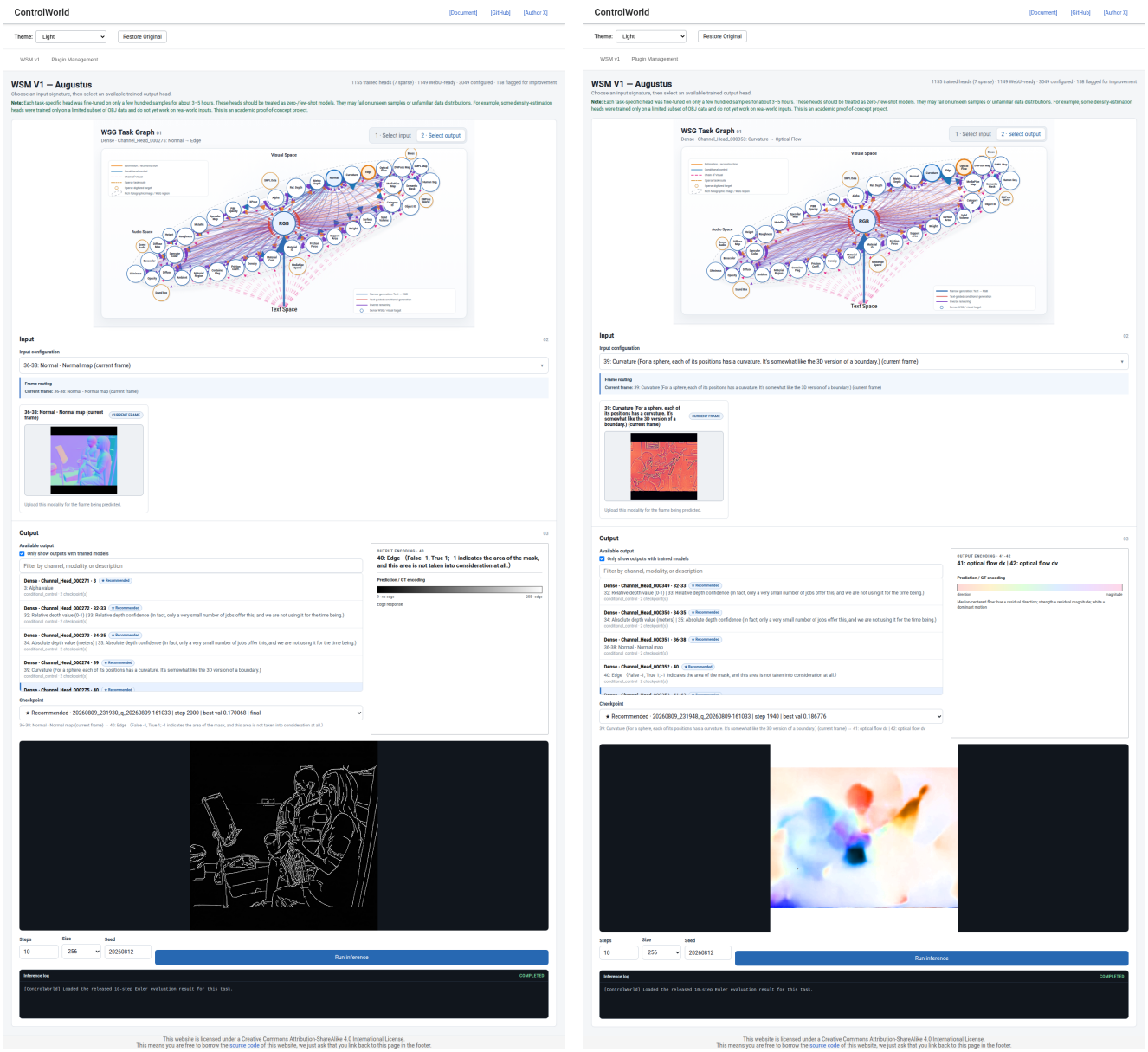


(a) **Head 000119:** RelDepth → Normal.



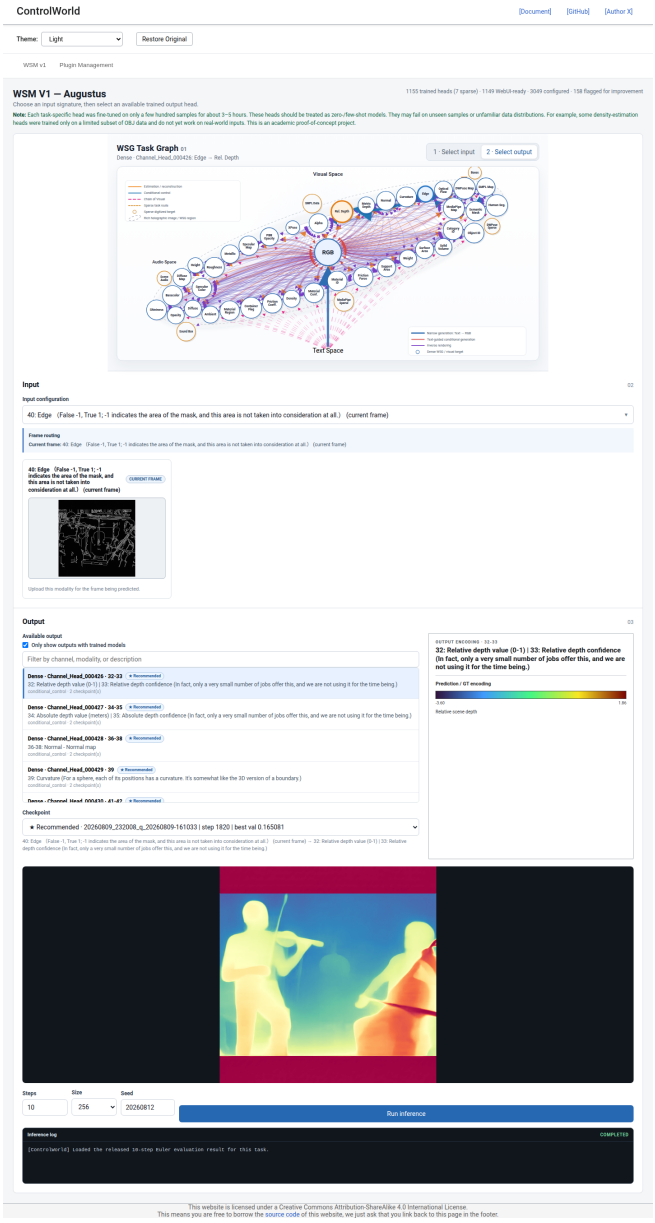
(b) **Head 000197:** AbsDepth → Curve.

Figure 60. ControlWorld desktop workflows for Qualitative Part IX: Head 000119 (RelDepth to Normal) and Head 000197 (AbsDepth to Curve). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

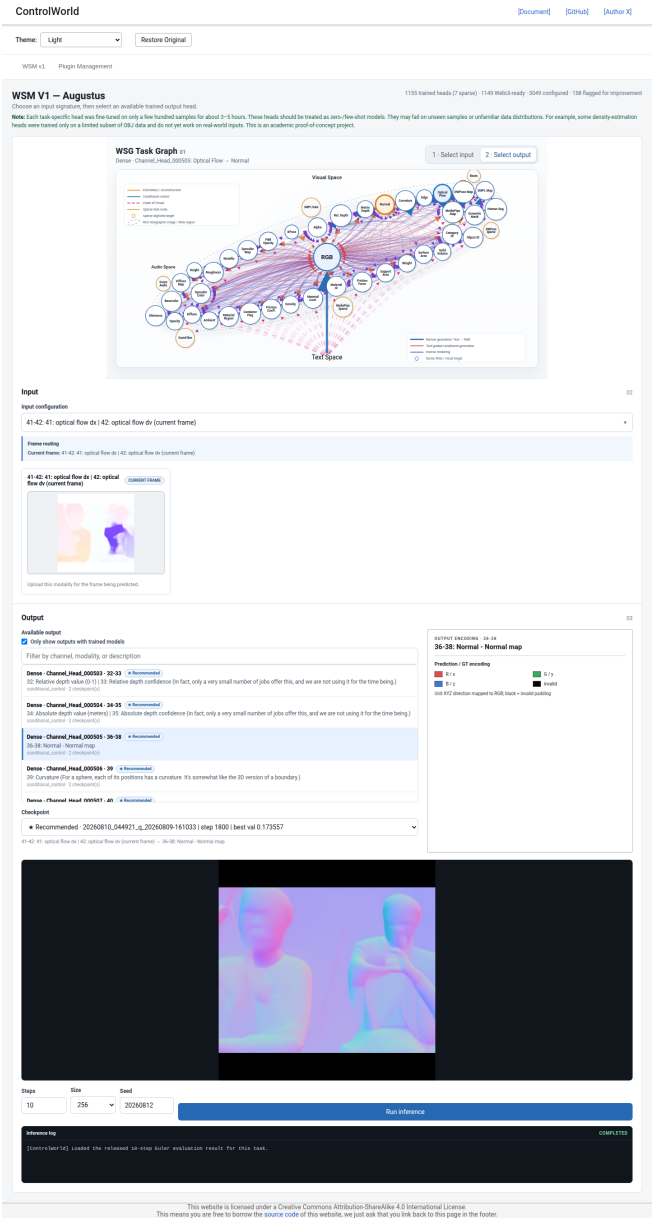


(a) **Head 000275:** Normal → Edge. (b) **Head 000353:** Curve → Flow.

Figure 61. ControlWorld desktop workflows for Qualitative Part IX: Head 000275 (Normal to Edge) and Head 000353 (Curve to Flow). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

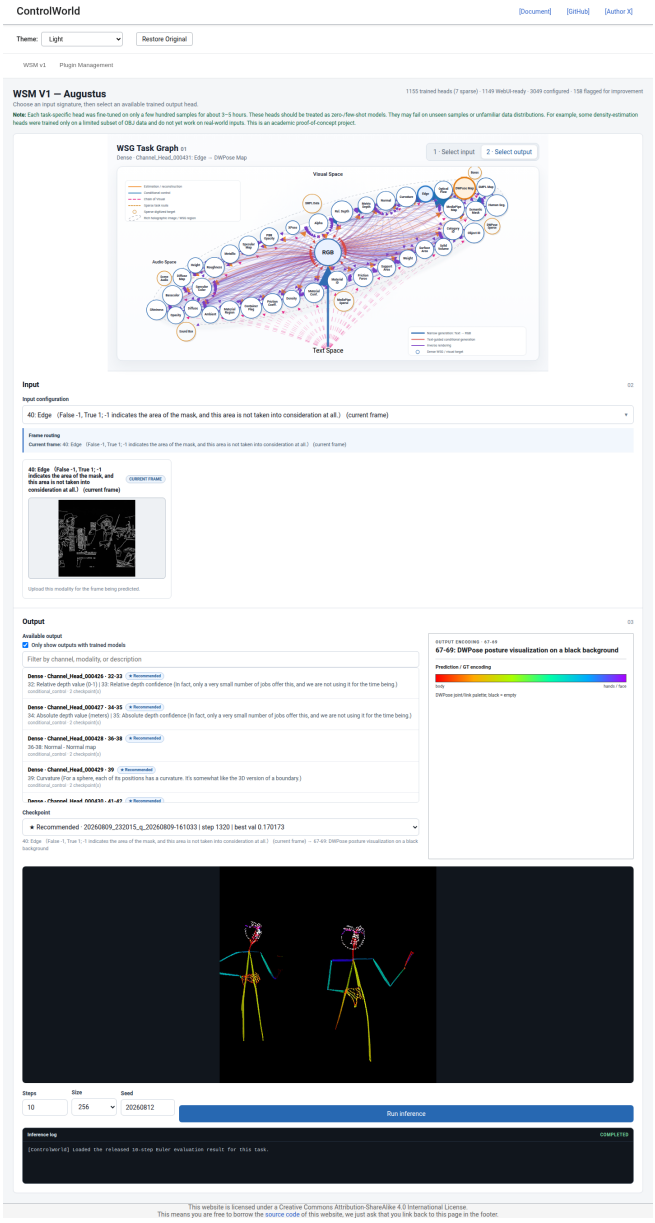


(a) **Head 000426:** Edge → RelDepth.

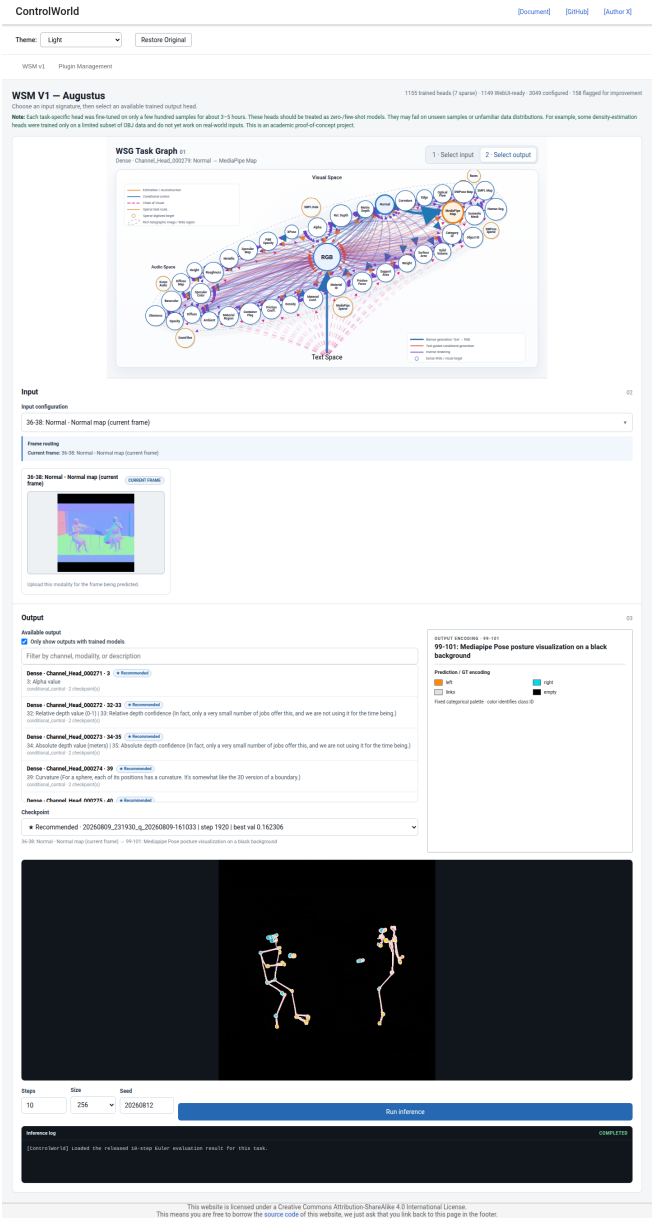


(b) **Head 000505:** Flow → Normal.

Figure 62. ControlWorld desktop workflows for Qualitative Part IX: Head 000426 (Edge to RelDepth) and Head 000505 (Flow to Normal). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

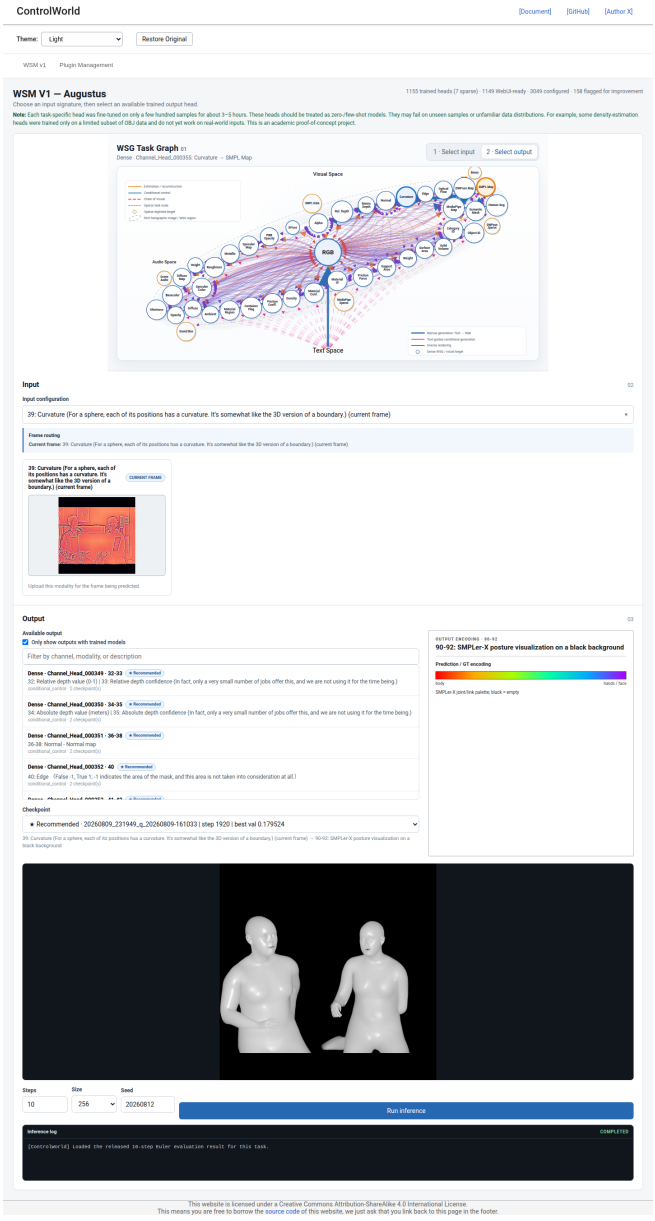


(a) **Head 000431:** Edge → DWPose.

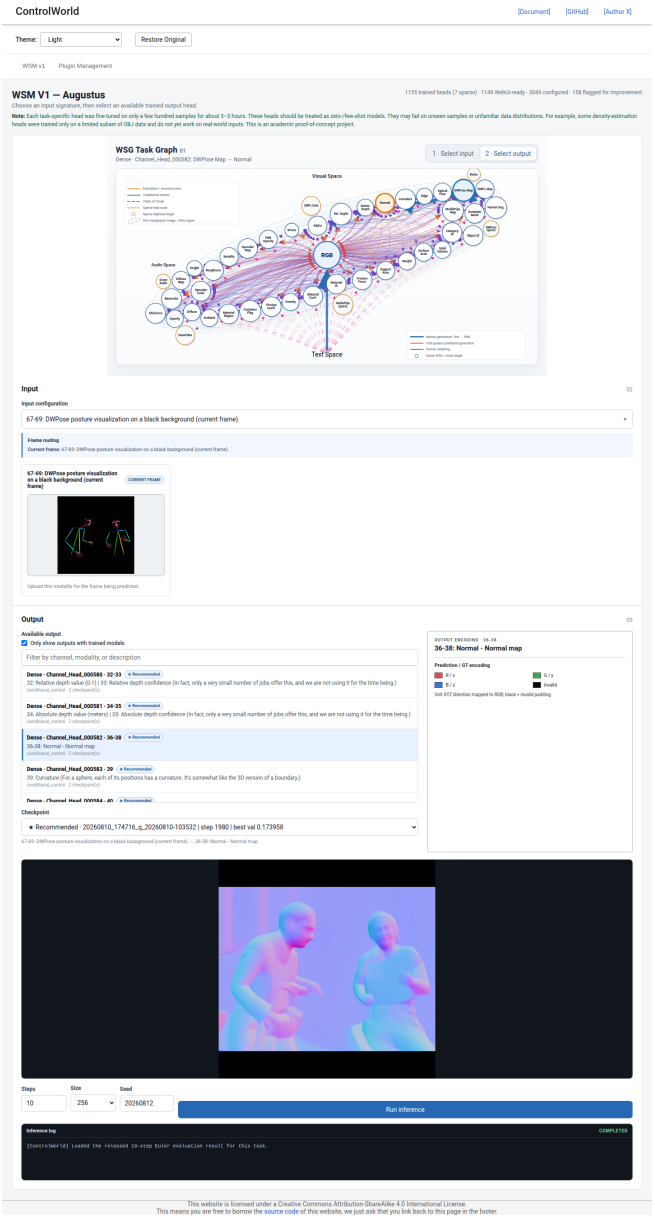


(b) **Head 000279:** Normal → MPPose.

Figure 63. ControlWorld desktop workflows for Qualitative Part IX: Head 000431 (Edge to DWPose) and Head 000279 (Normal to MPPose). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

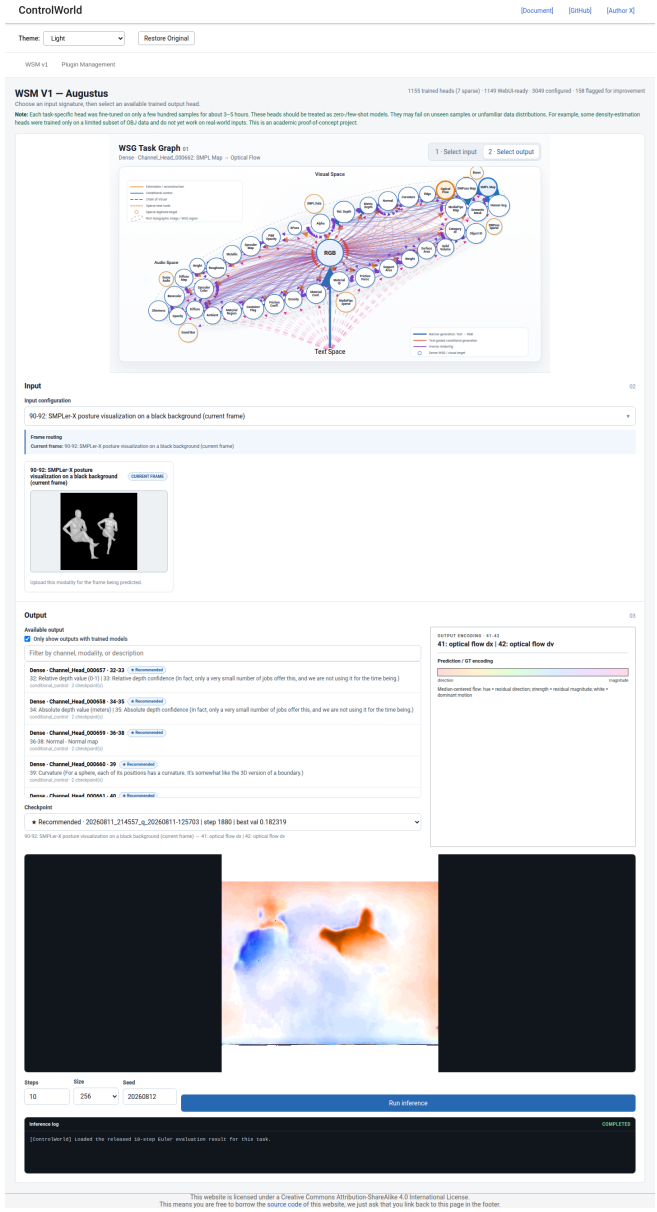


(a) **Head 000355:** Curve → SMPL-X.

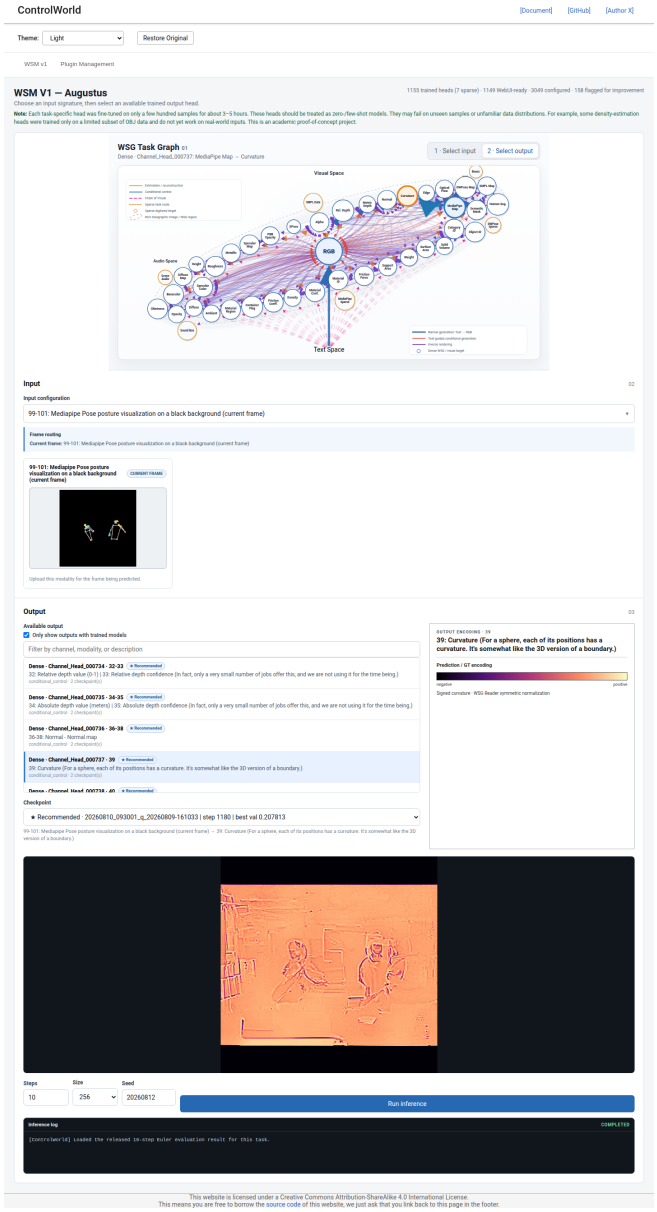


(b) **Head 000582:** DWPose → Normal.

Figure 64. ControlWorld desktop workflows for Qualitative Part IX: Head 000355 (Curve to SMPL-X) and Head 000582 (DWPose to Normal). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

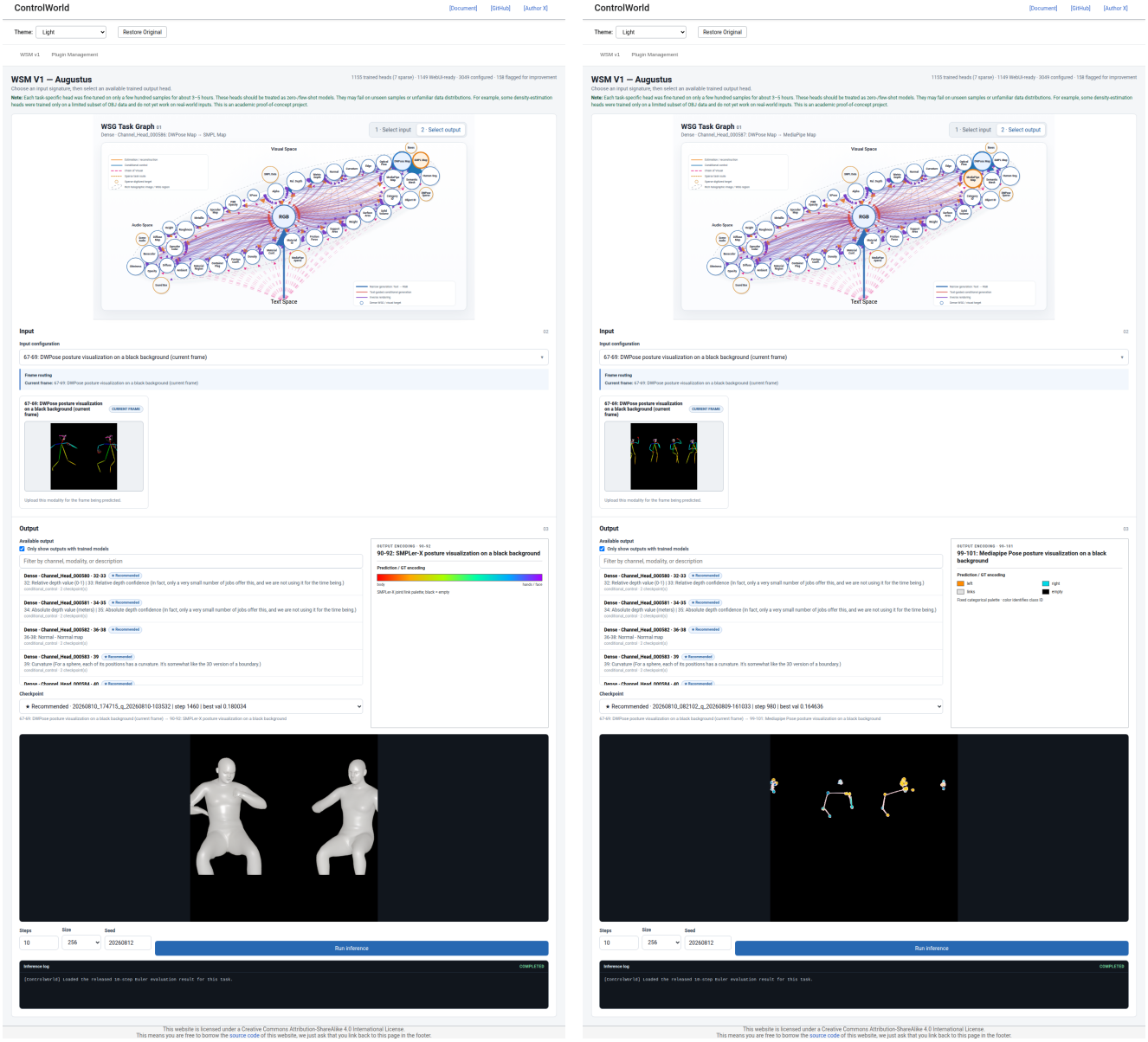


(a) **Head 000662:** SMPL-X → Flow.



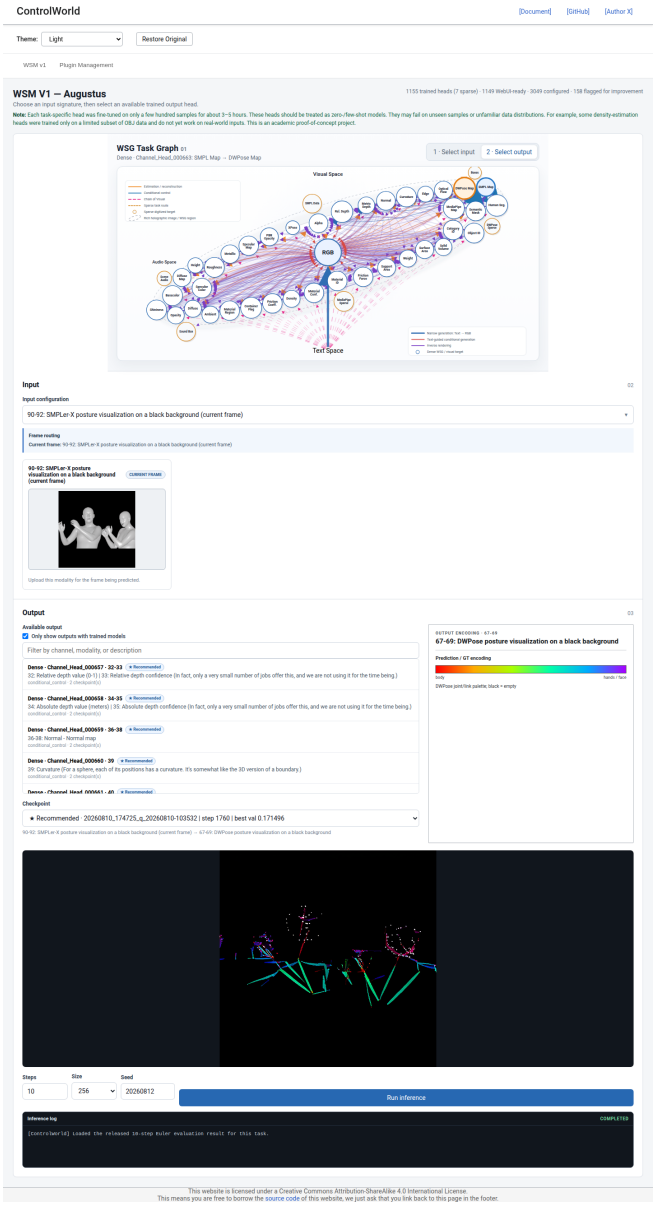
(b) **Head 000737:** MPPose → Curve.

Figure 65. ControlWorld desktop workflows for Qualitative Part IX: Head 000662 (SMPL-X to Flow) and Head 000737 (MPPose to Curve). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

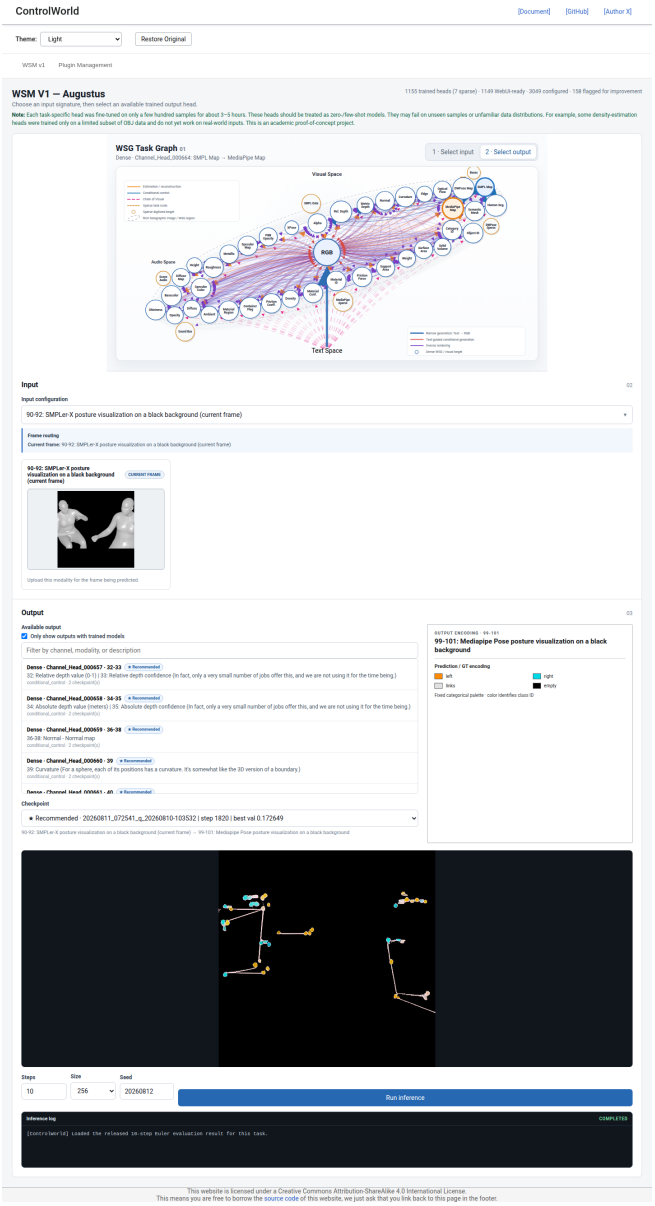


(a) **Head 000586:** DWPose → SMPL-X. (b) **Head 000587:** DWPose → MPPose.

Figure 66. ControlWorld desktop workflows for Qualitative Part IX: Head 000586 (DWPose to SMPL-X) and Head 000587 (DWPose to MPPose). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

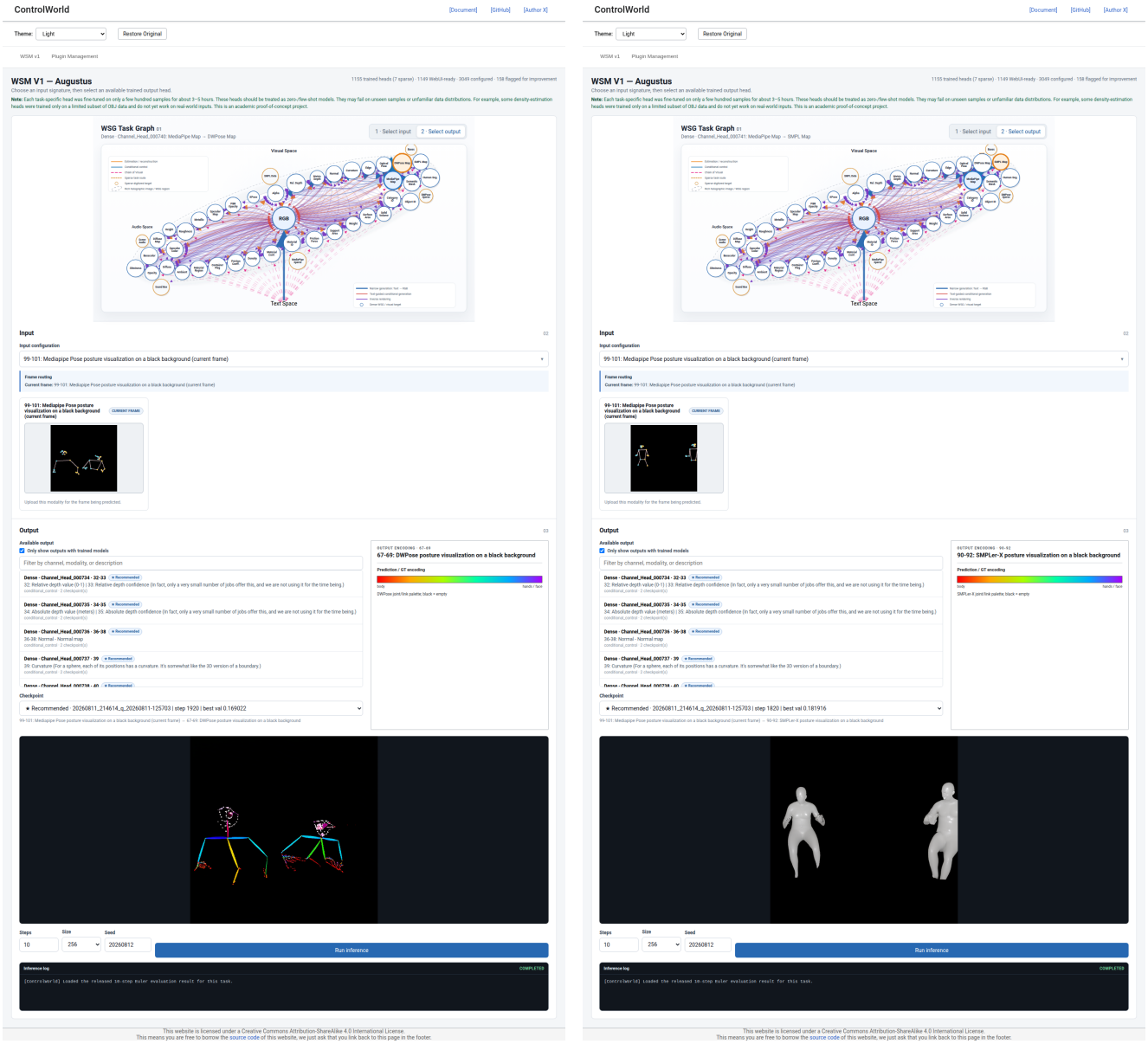


(a) **Head 000663:** SMPL-X → DWPose.



(b) **Head 000664:** SMPL-X → MPPose.

Figure 67. ControlWorld desktop workflows for Qualitative Part IX: Head 000663 (SMPL-X to DWPose) and Head 000664 (SMPL-X to MPPose). Each panel preserves the full task graph, input, sampling, and decoded-output interface.



(a) **Head 000740:** MPPose → DWPose. (b) **Head 000741:** MPPose → SMPL-X.

Figure 68. ControlWorld desktop workflows for Qualitative Part IX: Head 000740 (MPPose to DWPose) and Head 000741 (MPPose to SMPL-X). Each panel preserves the full task graph, input, sampling, and decoded-output interface.

12. Appendix: Historical Task Interface Analysis

To compare WSM with the tasks studied in previous Computer Vision research, we manually reviewed 1,947 task entries collected from the final Papers with Code archive [65]. For each task, we read its name and description and checked the listed categories, benchmarks, and metrics. We count a task as visual when visual data appears in its input or output, or when visual evidence is necessary to solve it. This includes vision–language and audiovisual tasks, but excludes entries based only on text, audio, graphs, or physiological signals.

The analysis separates two questions. The A–D assessment asks whether each historical visual task can be represented by the WSM dense or sparse interface system; all 1,477 visually grounded entries receive such a route. The second assessment asks whether each of 1,072 WSM cross-variable tasks had already been defined with the same input evidence and output target. It measures new task creation and does not measure compatibility with the historical catalog.

We organize the analysis into four ordered difficulty levels, from **A** (hardest) to **D** (easiest). **A** requires additional structured, temporal, interactive, or language-output development, for which WSM's sparse mode is the intended extension mechanism. **B** requires a minor interface, loss, label, data-adapter, or decoder adjustment. **C** keeps the method unchanged but requires a new aligned modality or dataset and corresponding training. **D** is directly compatible with an existing interface and represents an application, protocol, style, or preference scenario. The easiest D-level cases can be executed by the existing model without changing the loss or introducing a new task-specific training set. **X** marks an archive cross-listing with no visual role and is excluded from the difficulty ordering. Thus, $\mathrm{A} > \mathrm{B} > \mathrm{C} > \mathrm{D}$ orders the tasks by the engineering effort needed to realize them in WSM. Table 13 summarizes this review.

Among the 1,477 visual or vision-related tasks, 231 already fit an existing WSM interface, 255 require a new aligned modality or dataset, 286 require a minor interface or output adjustment, and 705 require a structured sparse extension. The central result is that every reviewed visual task has a route through the dense or sparse WSM interface system; the four levels describe how much additional work that route requires.

12.1. Novelty of evaluated WSM tasks

We separately examine how many of the 1,072 evaluated cross-variable tasks introduce an input–output task absent from the historical literature. Of these tasks, 655 have no historical exact task match, while 401 remain conservative ambiguities because the published task name does not specify its input evidence, and 16 have a clear match with the same input and output. The first count is the strict lower bound on new WSM tasks; adding the underspecified cases gives the upper estimate of 1,056. Table 14 summarizes these results.

Only 16 of the 1,072 mappings have a clear historical counterpart. At least 655 mappings are therefore new, and the estimate rises to 1,056 when the historically underspecified cases are included. Relative to the 1,477 established visual tasks, this expands the Computer Vision task space by 44.3–71.5%. Together, Tables 13 and 14 show that WSM both provides a common route for the established field and opens a substantially larger visual-task space.

Table 13. **Manual difficulty assessment of the final archived Computer Vision task inventory.** Counts come from manual review of 1,947 archive entries. The 470 X entries do not involve vision and are excluded from the visual-task percentages.

Difficulty assessment	Tasks	Archive share
Difficulty A: structured development via WSM sparse mode	705	36.2%
Difficulty B: minor interface, loss, label, or decoder adaptation	286	14.7%
Difficulty C: aligned-modality data and training	255	13.1%
Difficulty D: directly compatible application or scenario	231	11.9%
Non-visual archive cross-listing	470	24.1%
Total archived CV entries	1,947	100.0%

Table 14. **Historical overlap of the 1,072 evaluated WSM cross-variable tasks.** Rows group mappings according to whether the same input–output task appears in the historical literature.

Historical-overlap judgment	Tasks	Share of 1,072
No historical full-contract match (unambiguously new)	655	61.1%
Historical umbrella is underspecified (conservative ambiguity)	401	37.4%
Clear historical full-contract overlap	16	1.5%

12.2. Historical Tasks Covered by the Four Interfaces

Table 15 provides the task-level evidence behind the aggregate result in Table 13. Each row records the archived task definition and our judgment of how it relates to WSM. An A–D label identifies the route required to express a visual task through the dense or sparse WSM interface system, while X marks an entry that does not involve vision. All 1,477 visual or vision-related tasks receive an A–D route, and the remaining 470 nonvisual cross-listings receive X. The table therefore supports the central result of Table 13: the WSM interface system covers the complete reviewed visual-task set.

Table 15. **Task-level evidence for historical visual-task coverage.** Each row reports an archived task definition and the manual A–D/X decision used to derive Table 13. All 1,477 visual or vision-related entries receive an A–D route through the WSM interface system.

Archived task	Level	Archived description	Manual interface judgment
Semantic Segmentation	B	N/A	Semantic Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Object Detection	A	N/A	Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Image Classification	B	Image Classification is a fundamental task in vision recognition that aims to understand and categorize an image as a whole under a specific label. Unlike object detection	Image Classification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Visual Question Answering (VQA)	A	Visual Question Answering (VQA) is a task in computer vision that involves answering questions about an image. The goal of VQA is to teach machines to understand the content of an image and answer...	Visual Question Answering (VQA) requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Named Entity Recognition (NER)	X	Named Entity Recognition (NER) is a task of Natural Language Processing (NLP) that involves identifying and classifying named entities in a text into predefined categories such as person names, or...	Named Entity Recognition (NER) is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Pose Estimation	B	Pose Estimation is a computer vision task where the goal is to detect the position and orientation of a person or an object. Usually, this is done by predicting the location of specific keypoints ...	Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Anomaly Detection	D	Anomaly Detection is a binary classification identifying unusual or unexpected patterns in a dataset, which deviate significantly from the majority of the data. The goal of anomaly detection is to...	Anomaly Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
Action Recognition	A	Action Recognition is a computer vision task that involves recognizing human actions in videos or images. The goal is to classify and categorize the actions being performed in the video or image i...	Action Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Instance Segmentation	A	Instance Segmentation is a computer vision task that involves identifying and separating individual objects within an image, including detecting the boundaries of each object and assigning a uniqu...	Instance Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Speech Recognition	X	Speech Recognition is the task of converting spoken language into text. It involves recognizing the words spoken in an audio recording and transcribing them into a written format. The goal is to a...	Speech Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Information Retrieval	X	Information retrieval is the task of ranking a list of documents or search results in response to a query (Image credit: sudhanshumittal	Information Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
2D Object Detection	A	N/A	2D Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image Retrieval	A	Image Retrieval is a fundamental and long-standing computer vision task that involves finding images similar to a given query from a large database. It is often considered a form of fine-grained, ...	Image Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
2D Semantic Segmentation	B	N/A	2D Semantic Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Relation Extraction	A	Relation Extraction is the task of predicting attributes and relations for entities in a sentence. For example, given a sentence “Barack Obama was born in Honolulu, Hawaii.”, a relation classifier...	Relation Extraction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Image Generation	B	Image Generation (synthesis) is the task of generating new images from an existing dataset. - Unconditional generation refers to generating samples unconditionally from the dataset, i.e. $p...	Image Generation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Captioning	A	Image Captioning is the task of describing the content of an image in words. This task lies at the intersection of computer vision and natural language processing. Most image captioning systems us...	Image Captioning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Depth Estimation	B	Depth Estimation is the task of measuring the distance of each pixel relative to the camera. Depth is extracted from either monocular (single) or stereo (multiple views of a scene) images. Traditi...	Depth Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Autonomous Driving	D	Autonomous driving is the task of driving a vehicle without human conduction. Many of the state-of-the-art results can be found at more general task pages such as 3D Object Detection	Autonomous Driving adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
Object Tracking	A	Object tracking is the task of taking an initial set of object detections, creating a unique ID for each of the initial detections, and then tracking each of the objects as they move around frames...	Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Face Recognition	B	Facial Recognition is the task of making a positive identification of a face in a photo or video image against a pre-existing database of faces. It begins with detection - distinguishing human fac...	Face Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
3D Object Detection	A	3D Object Detection is a task in computer vision where the goal is to identify and locate objects in a 3D environment based on their shape, location, and orientation. It involves detecting the pre...	3D Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Link Prediction	X	Link Prediction is a task in graph and network analysis where the goal is to predict missing or future connections between nodes in a network. Given a partially observed network, the goal of link ...	Link Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
3D Reconstruction	A	3D Reconstruction is the task of creating a 3D model or representation of an object or scene from 2D images or other data sources. The goal of 3D reconstruction is to create a virtual representati...	3D Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Video Understanding	A	A crucial task of Video Understanding is to recognise and localise (in space and time) different actions or events appearing in the video. Source: Action Det...	Video Understanding requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Optical Character Recognition (OCR)	A	Optical Character Recognition or Optical Character Reader (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether fr...	Optical Character Recognition (OCR) requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Abstractive Text Summarization	X	Abstractive Text Summarization is the task of generating a short and concise summary that captures the salient ideas of the source text. The generated summaries potentially contain new phrases and...	Abstractive Text Summarization is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Medical Image Segmentation	C	Medical Image Segmentation is a computer vision task that involves dividing an medical image into multiple segments, where each segment represents a different object or structure of interest in th...	Medical Image Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Emotion Recognition	C	Emotion Recognition is an important area of research to enable effective human-computer interaction. Human emotions can be detected using speech signal, facial expressions, body language, and elec...	Emotion Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
3D Human Pose Estimation	B	3D Human Pose Estimation is a computer vision task that involves estimating the 3D positions and orientations of body joints and bones from 2D images or videos. The goal is to reconstruct the 3D p...	3D Human Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...

Archived task	Level	Archived description	Manual interface judgment
Hate Speech Detection	X	Hate speech detection is the task of detecting if communication such as text, audio, and so on contains hatred and or encourages violence towards a person or a group of people. This is usually based o...	Hate Speech Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Novel View Synthesis	A	Synthesize a target image with an arbitrary target camera pose from given source images and their camera poses. See Wiki for more introductions. T...	Novel View Synthesis requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Object Recognition	B	Object recognition is a computer vision technique for detecting + classifying objects in images or videos. Since this is a combined task of object detection plus image classification, the state-of-the...	Object Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Semantic Parsing	X	Semantic Parsing is the task of transducing natural language utterances into formal meaning representations. The target meaning representations can be defined according to a wide variety of formal...	Semantic Parsing is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Segmentation	B	N/A	Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Image Clustering	A	Models that partition the dataset into semantically meaningful clusters without having access to the ground truth labels. Image credit: ImageNet clustering ...	Image Clustering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Visual Reasoning	A	Ability to understand actions and reasoning associated with any visual images	Visual Reasoning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Image Super-Resolution	B	Image Super-Resolution is a machine learning task where the goal is to increase the resolution of an image, often by a factor of 4x or more, while maintaining its content and details as much as po...	Image Super-Resolution is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Segmentation	B	Image Segmentation is a computer vision task that involves dividing an image into multiple segments or regions, each of which corresponds to a different object or part of an object. The goal of im...	Image Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Video Question Answering	A	N/A	Video Question Answering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Multi-Object Tracking	A	Multi-Object Tracking is a task in computer vision that involves detecting and tracking multiple objects within a video sequence. The goal is to identify and locate objects of interest in each fra...	Multi-Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Scene Understanding	B	Scene understanding involves interpreting the visual information of a scene, including objects, their spatial relationships, and the overall layout. It goes beyond simple object recognition by conside...	Scene Understanding is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Temporal Action Localization	A	Temporal Action Localization aims to detect activities in the video stream and output beginning and end timestamps. It is closely related to Temporal Action Proposal Generation.	Temporal Action Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Fine-Grained Image Classification	B	Fine-Grained Image Classification is a task in computer vision where the goal is to classify images into subcategories within a larger category. For example, classifying different species of birds...	Fine-Grained Image Classification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
3D Semantic Segmentation	C	3D Semantic Segmentation is a computer vision task that involves dividing a 3D point cloud or 3D mesh into semantically meaningful parts or regions. The goal of 3D semantic segmentation is to iden...	3D Semantic Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Face Detection	A	Face Detection is a computer vision task that involves automatically identifying and locating human faces within digital images or videos. It is a fundamental technology that underpins many applic...	Face Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Retrieval	A	A methodology that involves selecting relevant data or examples from a large dataset to support tasks like prediction, learning, or inference. It enhances models by providing context or additional inf...	Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured logic;...

Archived task	Level	Archived description	Manual interface judgment
Video Captioning	A	Video Captioning is a task of automatic captioning a video by understanding the action and event in the video which can help in the retrieval of the video efficiently through text. ¡span clas...	Video Captioning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Panoptic Segmentation	A	Panoptic Segmentation is a computer vision task that combines semantic segmentation and instance segmentation to provide a comprehensive understanding of the scene. The goal of panoptic segmentati...	Panoptic Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Text Retrieval	A	N/A	Text Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
3D Pose Estimation	B	Image credit: GSNet: Joint Vehicle Pose and Shape Reconstruction with Geometrical and Scene-aware Supervision , ECCV'20	3D Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Optical Flow Estimation	B	Optical Flow Estimation is a computer vision task that involves computing the motion of objects in an image or a video sequence. The goal of optical flow estimation is to determine the movement of...	Optical Flow Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Video Retrieval	A	The objective of video retrieval is as follows: given a text query and a pool of candidate videos, select the video which corresponds to the text query. Typically, the videos are returned as a ranked...	Video Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Trajectory Prediction	A	Trajectory Prediction is the problem of predicting the short-term (1-3 seconds) and long-term (3-5 seconds) spatial coordinates of various road-agents such as cars, buses, pedestrians, rickshaws, ...	Trajectory Prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Stance Detection	X	Stance detection is the extraction of a subject's reaction to a claim made by a primary actor. It is a core part of a set of approaches to fake news assessment. Example: * Source: "Apples are the mo...	Stance Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Action Detection	A	Action Detection aims to find both where and when an action occurs within a video clip and classify what the action is taking place. Typically results are given in the form of action tublets, which ar...	Action Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Automatic Speech Recognition (ASR)	X	Automatic Speech Recognition (ASR) involves converting spoken language into written text. It is designed to transcribe spoken words into text in real-time, allowing people to communicate with comp...	Automatic Speech Recognition (ASR) is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Monocular Depth Estimation	B	N/A	Monocular Depth Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Action Classification	A	N/A	Action Classification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Visual Question Answering	A	MLLM Leaderboard	Visual Question Answering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image-to-Image Translation	B	Image-to-Image Translation is a task in computer vision and machine learning where the goal is to learn a mapping between an input image and an output image, such that the output image can be used...	Image-to-Image Translation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Skeleton Based Action Recognition	A	N/A	Skeleton Based Action Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Activity Recognition	A	Human Activity Recognition is the problem of identifying events performed by humans given a video input. It is formulated as a binary (or multiclass) classification problem of outputting activity ...	Activity Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Fake News Detection	A	Fake News Detection is a natural language processing task that involves identifying and classifying news articles or other types of text as real or fake. The goal of fake news detection is to deve...	Fake News Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Sign Language Recognition	A	Sign Language Recognition is a computer vision and natural language processing task that involves automatically recognizing and translating sign language gestures into written or spoken language. ...	Sign Language Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Visual Object Tracking	A	Visual Object Tracking is an important research topic in computer vision, image understanding and pattern recognition. Given the initial state (centre location and scale) of a target in the first ...	Visual Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Facial Expression Recognition (FER)	B	Facial Expression Recognition (FER) is a computer vision task aimed at identifying and categorizing emotional expressions depicted on a human face. The goal is to automate the process of determini...	Facial Expression Recognition (FER) is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Scene Text Recognition	A	See Scene Text Detection for leaderboards in this task.	Scene Text Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Autonomous Vehicles	D	Autonomous vehicles is the task of making a vehicle that can guide itself without human conduction. Many of the state-of-the-art results can be found at more general task pages such as 3D Object D...	Autonomous Vehicles adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant...
Unsupervised Anomaly Detection	D	The objective of Unsupervised Anomaly Detection is to detect previously unseen rare objects or events without any prior knowledge about these. The only information available is that the percentage...	Unsupervised Anomaly Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Object Counting	B	The goal of Object Counting task is to count the number of object instances in a single image or video sequence. It has many real-world applications such as traffic flow monitoring, crowdedness es...	Object Counting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Visual Place Recognition	A	Visual Place Recognition is the task of matching a view of a place with a different view of the same place taken at a different time. Source: Visual place reco...	Visual Place Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Visual Tracking	A	Visual Tracking is an essential and actively researched problem in the field of computer vision with various real-world applications such as robotic services, smart surveillance systems, autonomou...	Visual Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Neural Architecture Search	X	Neural architecture search (NAS) is a technique for automating the design of artificial neural networks (ANN), a widely used model in the field of machine learning. NAS essentially takes the proce...	Neural Architecture Search is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
2D Human Pose Estimation	B	What is Human Pose Estimation? Human pose estimation is the process of estimating the configuration of the body (pose) from a single, typically monocular, image. Background. Human pose estimation is ...	2D Human Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Emotion Classification	A	Emotion classification, or emotion categorization, is the task of recognising emotions to classify them into the corresponding category. Given an input, classify it as 'neutral or no emotion' or as on...	Emotion Classification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Visual Localization	A	Visual Localization is the problem of estimating the camera pose of a given image relative to a visual representation of a known scene. Source: Fine-Grained Seg...	Visual Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Visual Odometry	A	Visual Odometry is an important area of information fusion in which the central aim is to estimate the pose of a robot using data collected by visual sensors. So...	Visual Odometry requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Intent Detection	X	Intent Detection is a task of determining the underlying purpose or goal behind a user's search query given a context. The task plays a significant role in search and recommendations. A traditio...	Intent Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
DeepFake Detection	A	DeepFake Detection is the task of detecting fake videos or images that have been generated using deep learning techniques. Deepfakes are created by using machine learning algorithms to manipulate ...	DeepFake Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Music Information Retrieval	X	N/A	Music Information Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Video Prediction	A	Script for Amee Marketing & Trading Company Short Video *(Duration: 45-60 seconds)* — Opening Scene (0:00-0:05): - *Visual:* Close-up of fresh organic grains spilling gently into a woo...	Video Prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...

Archived task	Level	Archived description	Manual interface judgment
Text-to-Image Generation	A	The development of the brain's blood supply in an embryo involves a complex process with several stages. Initially, there is a network of connections between the carotid and basilar artery systems tha...	Text-to-Image Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Few-Shot Image Classification	D	Few-Shot Image Classification is a computer vision task that involves training machine learning models to classify images into predefined categories using only a few labeled examples of each categ...	Few-Shot Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Face Verification	B	Face Verification is a machine learning task in computer vision that involves determining whether two facial images belong to the same person or not. The task involves extracting features from the...	Face Verification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Out-of-Distribution Detection	D	Detect out-of-distribution or anomalous examples.	Out-of-Distribution Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Cross-Modal Retrieval	A	Cross-Modal Retrieval (CMR) is a task of retrieving items across different modalities, such as image, text, video, and audio. The core challenge of CMR is the *heterogeneity gap*, which arises bec...	Cross-Modal Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Generation	A	(Various Video Generation Tasks. Gif credit: MaGViT)	Video Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Speech Enhancement	A	Speech Enhancement is a signal processing task that involves improving the quality of speech signals captured under noisy or degraded conditions. The goal of speech enhancement is to make speech s...	Speech Enhancement requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Crowd Counting	B	N/A	Crowd Counting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Human-Object Interaction Detection	A	Human-Object Interaction (HOI) detection is a task of identifying "a set of interactions" in an image, which involves the i) localization of the subject (i.e., humans) and target (i.e., objects) of in...	Human-Object Interaction Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Denoising	B	Denoising is a task in image processing and computer vision that aims to remove or reduce noise from an image. Noise can be introduced into an image due to various reasons, such as camera sensor l...	Denoising is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment are...
Low-Light Image Enhancement	B	N/A	Low-Light Image Enhancement is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
6D Pose Estimation	B	Image: Zeng et al	6D Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Hand Pose Estimation	B	Hand pose estimation is the task of finding the joints of the hand from an image or set of video frames. (Image credit: Pose-REN	Hand Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Scene Classification	B	Scene Classification is a task in which scenes from photographs are categorically classified. Unlike object classification, which focuses on classifying prominent objects in the foreground, Scene ...	Scene Classification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Automatic Speech Recognition	X	N/A	Automatic Speech Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Speech Emotion Recognition	A	Speech Emotion Recognition is a task of speech processing and computational paralinguistics that aims to recognize and categorize the emotions expressed in spoken language. The goal is to determin...	Speech Emotion Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Classification	A	Video Classification is the task of producing a label that is relevant to the video given its frames. A good video level classifier is one that not only provides accurate frame labels, but also be...	Video Classification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image Denoising	B	Image Denoising is a computer vision task that involves removing noise from an image. Noise can be introduced into an image during acquisition or processing, and can reduce image quality and make ...	Image Denoising is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Cell Segmentation	C	Cell Segmentation is a task of splitting a microscopic image domain into segments, which represent individual instances of cells. It is a fundamental step in many biomedical studies, and it is reg...	Cell Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...

Archived task	Level	Archived description	Manual interface judgment
Speech Synthesis	A	Speech synthesis is the task of generating speech from some other modality like text, lip movements etc. Please note that the leaderboards here are not really comparable between studies - as they ...	Speech Synthesis requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Super-Resolution	B	Super-Resolution is a task in computer vision that involves increasing the resolution of an image or video by generating missing high-frequency details from low-resolution input. The goal is to pr...	Super-Resolution is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Handwriting Recognition	A	Image source: Handwriting Recognition of Historical Documents with few labeled data	Handwriting Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Molecular Property Prediction	X	Molecular property prediction is the task of predicting the properties of a molecule from its structure.	Molecular Property Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Simultaneous Localization and Mapping	A	Simultaneous localization and mapping (SLAM) is the task of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it. ¡span style...	Simultaneous Localization and Mapping requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Medical Diagnosis	C	Medical Diagnosis is the process of identifying the disease a patient is affected by, based on the assessment of specific risk factors, signs, symptoms and results of exams. ¡span class="desc...	Medical Diagnosis is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Sound Event Detection	X	Sound Event Detection (SED) is the task of recognizing the sound events and their respective temporal start and end time in a recording. Sound events in real life do not always occur in isolation,...	Sound Event Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Traffic Prediction	X	Traffic Prediction is a task that involves forecasting traffic conditions, such as the volume of vehicles and travel time, in a specific area or along a particular road. This task is important for...	Traffic Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Aspect-Based Sentiment Analysis (ABSA)	X	Aspect-Based Sentiment Analysis (ABSA) is a Natural Language Processing task that aims to identify and extract the sentiment of specific aspects or components of a product or service. ABSA typical...	Aspect-Based Sentiment Analysis (ABSA) is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Named Entity Recognition	X	N/A	Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Change Detection	A	Change Detection is a computer vision task that involves detecting changes in an image or video sequence over time. The goal is to identify areas in the image or video that have undergone changes,...	Change Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Stereo Matching	B	Stereo Matching is one of the core technologies in computer vision, which recovers 3D structures of real world from 2D images. It has been widely used in areas such as autonomous driving, augmente...	Stereo Matching is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
3D Hand Pose Estimation	B	Image: Zimmerman et l	3D Hand Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Manipulation Detection	B	The task of detecting images or image parts that have been tampered or manipulated (sometimes also referred to as doctored). This typically encompasses image splicing, copy-move, or image inpainting.	Image Manipulation Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Semantic Textual Similarity	X	Semantic textual similarity deals with determining how similar two pieces of texts are. This can take the form of assigning a score from 1 to 5. Related tasks are paraphrase or duplicate identificati...	Semantic Textual Similarity is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Medical Image Classification	C	Medical Image Classification is a task in medical image analysis that involves classifying medical images, such as X-rays, MRI scans, and CT scans, into different categories based on the type of i...	Medical Image Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Video Object Segmentation	A	Video object segmentation is a binary labeling problem aiming to separate foreground object(s) from the background region of a video. For leaderboards please refer to the different subtasks.	Video Object Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Visual Navigation	A	Visual Navigation is the problem of navigating an agent, e.g. a mobile robot, in an environment using camera input only. The agent is given a target image (an image it will see from the target pos...	Visual Navigation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Animal Pose Estimation	B	Animal pose estimation is the task of identifying the pose of an animal. (Image credit: Using DeepLabCut for 3D markerless pose estimation across species an...	Animal Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Robot Navigation	A	The fundamental objective of mobile Robot Navigation is to arrive at a goal position without collision. The mobile robot is supposed to be aware of obstacles and move freely in different working s...	Robot Navigation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Motion Synthesis	A	Creating a video where people in the images move (such as blinking or smiling) requires specialized AI technology like Deepfake or Motion Synthesis, which I cannot do directly here. However, if y...	Motion Synthesis requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Transductive Zero-Shot Classification	D	N/A	Transductive Zero-Shot Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Object Localization	A	Object Localization is the task of locating an instance of a particular object category in an image, typically by specifying a tightly cropped bounding box centered on the instance. An object prop...	Object Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Multiple Object Tracking	A	Multiple Object Tracking is the problem of automatically identifying multiple objects in a video and representing them as a set of trajectories with high accuracy. ¡span class=”description-sourc...	Multiple Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Gaze Estimation	B	Gaze Estimation is a task to predict where a person is looking at given the person's full face. The task contains two directions: 3-D gaze vector and 2-D gaze position estimation. 3-D gaze vector ...	Gaze Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Action Segmentation	A	Action Segmentation is a challenging problem in high-level video understanding. In its simplest form, Action Segmentation aims to segment a temporally untrimmed video by time and label each segmen...	Action Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
3D Instance Segmentation	A	Image: OccuSeg	3D Instance Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Text-To-Speech Synthesis	X	Text-To-Speech Synthesis is a machine learning task that involves converting written text into spoken words. The goal is to generate synthetic speech that sounds natural and resembles human speech...	Text-To-Speech Synthesis is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Zero-Shot Video Question Answer	A	N/A	Zero-Shot Video Question Answer requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Image Restoration	B	Image Restoration is a family of inverse problems for obtaining a high quality image from a corrupted input image. Corruption may occur due to the image-capture process (e.g., noise, lens blur), p...	Image Restoration is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Enhancement	B	Image Enhancement is basically improving the interpretability or perception of information in images for human viewers and providing 'better' input for other automated image processing techniques....	Image Enhancement is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Semantic Similarity	X	The main objective Semantic Similarity is to measure the distance between the semantic meanings of a pair of words, phrases, sentences, or documents. For example, the word “car” is more similar to...	Semantic Similarity is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Face Alignment	B	N/A	Face Alignment is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Image Inpainting	B	Remove the preview tag from template	Image Inpainting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Dehazing	B	(Image credit: Densely Connected Pyramid Dehazing Network)	Image Dehazing is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
3D Object Tracking	A	3D Object Tracking is a computer vision task dedicated to monitoring and precisely locating objects as they navigate within a three-dimensional environment. It frequently utilizes 3D object detection ...	3D Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...

Archived task	Level	Archived description	Manual interface judgment
Action Recognition In Videos	A	Action Recognition in Videos is a task in computer vision and pattern recognition where the goal is to identify and categorize human actions performed in a video sequence. The task involves analyz...	Action Recognition In Videos requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Scene Text Detection	A	Scene Text Detection is a computer vision task that involves automatically identifying and localizing text within natural images or videos. The goal of scene text detection is to develop algorithm...	Scene Text Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Anomaly Detection	A	N/A	Video Anomaly Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Lane Detection	B	Lane Detection is a computer vision task that involves identifying the boundaries of driving lanes in a video or image of a road scene. The goal is to accurately locate and track the lane markings...	Lane Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Face Anti-Spoofing	B	N/A	Face Anti-Spoofing is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Pedestrian Detection	A	Pedestrian detection is the task of detecting pedestrians from a camera. Further state-of-the-art results (e.g. on the KITTI dataset) can be found at 3D Object Detection	Pedestrian Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
RGB Salient Object Detection	B	RGB Salient object detection is a task-based on a visual attention mechanism, in which algorithms aim to explore objects or regions more attentive than the surrounding areas on the scene or RGB images...	RGB Salient Object Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Salient Object Detection	B	N/A	Salient Object Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Deblurring	B	Deblurring is a computer vision task that involves removing the blurring artifacts from images or videos to restore the original, sharp content. Blurring can be caused by various factors such as c...	Deblurring is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment are...
Lesion Segmentation	C	N/A	Lesion Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Fact Verification	A	Fact verification, also called "fact checking", is a process of verifying facts in natural text against a database of facts.	Fact Verification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Beat Tracking	X	Determine the positions of all beats in a music recording.	Beat Tracking is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Motor Imagery Decoding (left-hand vs right-hand)	C	N/A	Motor Imagery Decoding (left-hand vs right-hand) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
Image Quality Assessment	B	N/A	Image Quality Assessment is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Scene Recognition	B	N/A	Scene Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Facial Landmark Detection	B	Facial Landmark Detection is a computer vision task that involves detecting and localizing specific points or landmarks on a face, such as the eyes, nose, mouth, and chin. The goal is to accuratel...	Facial Landmark Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Grammatical Error Correction	X	Grammatical Error Correction (GEC) is the task of correcting different kinds of errors in text such as spelling, punctuation, grammatical, and word choice errors. GEC is typically formulated as a ...	Grammatical Error Correction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Video Summarization	A	Video Summarization aims to generate a short synopsis that summarizes the video content by selecting its most informative and important parts. The produced summary is usually composed of a set of ...	Video Summarization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Active Learning	D	Active Learning is a paradigm in supervised machine learning which uses fewer training examples to achieve better optimization by iteratively training a predictor, and using the predictor in each ...	Active Learning adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of an...
Joint Entity and Relation Extraction	X	N/A	Joint Entity and Relation Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Event Extraction	X	Determine the extent of the events in a text. Other names: Event Tagging; Event Identification	Event Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
3D Action Recognition	A	Image: Rahmani et al	3D Action Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Hand Gesture Recognition	A	N/A	Hand Gesture Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Trajectory Forecasting	A	N/A	Trajectory Forecasting requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Super-Resolution	A	Video Super-Resolution is a computer vision task that aims to increase the resolution of a video sequence, typically from lower to higher resolutions. The goal is to generate high-resolution video...	Video Super-Resolution requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Single-View 3D Reconstruction	A	N/A	Single-View 3D Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Frame Interpolation	A	The goal of Video Frame Interpolation is to synthesize several frames in the middle of two adjacent frames of the original video. Video Frame Interpolation can be applied to generate slow motion v...	Video Frame Interpolation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Human Activity Recognition	A	Classify various human activities	Human Activity Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Gesture Recognition	A	Gesture Recognition is an active field of research with applications such as automatic recognition of sign language, interaction of humans and robots or for new ways of controlling video games. ...	Gesture Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
2D Pose Estimation	B	detective pose	2D Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Computed Tomography (CT)	C	The term “computed tomography”, or CT, refers to a computerized x-ray imaging procedure in which a narrow beam of x-rays is aimed at a patient and quickly rotated around the body, producing signals th...	Computed Tomography (CT) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Saliency Detection	B	Saliency Detection is a preprocessing step in computer vision which aims at finding salient objects in an image. Source: An Unsupervised Game-Theoretic Approach...	Saliency Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Video Inpainting	A	The goal of Video Inpainting is to fill in missing regions of a given video sequence with contents that are both spatially and temporally coherent. Video Inpainting, also known as video completion...	Video Inpainting requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Sarcasm Detection	X	The goal of Sarcasm Detection is to determine whether a sentence is sarcastic or non-sarcastic. Sarcasm is a type of phenomenon with specific perlocutionary effects on the hearer, such as to break...	Sarcasm Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
License Plate Recognition	A	N/A	License Plate Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Tumor Segmentation	C	N/A	Tumor Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
3D Human Reconstruction	A	N/A	3D Human Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
3D Classification	C	N/A	3D Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Semi-Supervised Image Classification	D	Semi-supervised image classification leverages unlabelled data as well as labelled data to increase classification performance. You may want to read some blog posts to get an overview before readin...	Semi-Supervised Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Weakly Supervised Object Detection	A	N/A	Weakly Supervised Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Open Information Extraction	X	In natural language processing, open information extraction is the task of generating a structured, machine-readable representation of the information in text, usually in the form of triples or n-ary ...	Open Information Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Text-to-Video Generation	A	Ma grand-mère m’a raconté que quand elle était étudiante, elle avait un petit-ami. À l’âge de 18 ans, il a dû partir pour le service militaire, elle ne l’a pas attendu et elle a épousé quelqu’un d’aut...	Text-to-Video Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Emotion Recognition in Conversation	A	N/A	Emotion Recognition in Conversation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Boundary Detection	B	Boundary Detection is a vital part of extracting information encoded in images, allowing for the computation of quantities of interest including density, velocity, pressure, etc. ¡span class=...	Boundary Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Nested Named Entity Recognition	X	N/A	Nested Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Semantic Role Labeling	X	Semantic role labeling aims to model the predicate-argument structure of a sentence and is often described as answering "Who did what to whom". BIO notation is typically used for semantic role label...	Semantic Role Labeling is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Network Intrusion Detection	X	N/A	Network Intrusion Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
motion prediction	A	N/A	motion prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Handwritten Text Recognition	A	Handwritten Text Recognition (HTR) is the task of automatically identifying and transcribing handwritten text from images or scanned documents into machine-readable text. The goal is to develop a syst...	Handwritten Text Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Pose Tracking	A	Pose Tracking is the task of estimating multi-person human poses in videos and assigning unique instance IDs for each keypoint across frames. Accurate estimation of human keypoint-trajectories is ...	Pose Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Human Detection	A	DEFINICIÓN EL Granada SOUND es un festival de música india de ámbito nacional e internacional que se celebra en Granada, España. Desde su primera edición en 2012, el evento ha ido creciendo en popul...	Human Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Event Detection	A	N/A	Event Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Indoor Localization	A	Indoor localization is a fundamental problem in indoor location-based applications.	Indoor Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Instruction Following	A	Instruction following is the basic task of the model. This task is dedicated to evaluating the ability of the large model to follow human instructions. It is hoped that the model can generate controll...	Instruction Following requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Keypoint Detection	B	N/A	Keypoint Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Video Object Tracking	A	N/A	Video Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Dialogue State Tracking	X	N/A	Dialogue State Tracking is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Small Object Detection	A	Small Object Detection is a computer vision task that involves detecting and localizing small objects in images or videos. This task is challenging due to the small size and low resolution of the ...	Small Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Object Detection	A	N/A	Video Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Semi-Supervised Video Object Segmentation	A	The semi-supervised scenario assumes the user inputs a full mask of the object(s) of interest in the first frame of a video sequence. Methods have to produce the segmentation mask for that object(s) i...	Semi-Supervised Video Object Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Semi-Supervised Semantic Segmentation	D	N/A	Semi-Supervised Semantic Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Hyperspectral Image Classification	C	N/A	Hyperspectral Image Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Community Detection	X	Community Detection is one of the fundamental problems in network analysis, where the goal is to find groups of nodes that are, in some sense, more similar to each other than to the other nodes. ...	Community Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Downbeat Tracking	X	Determine the positions of all downbeats in a music recording.	Downbeat Tracking is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Vision and Language Navigation	A	N/A	Vision and Language Navigation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...

Archived task	Level	Archived description	Manual interface judgment
Multiple Instance Learning	A	Multiple Instance Learning is a type of weakly supervised learning algorithm where training data is arranged in bags, where each bag contains a set of instances $X=\\{x_1,x_2, \ldots,x_M\\}$, and ...	Multiple Instance Learning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image Registration	B	Image registration is the process of transforming different sets of data into one coordinate system. Data may be multiple photographs, data from different sensors, times, depths, or viewpoints. It is ...	Image Registration is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Reconstruction	B	N/A	Image Reconstruction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Compression	B	Image Compression is an application of data compression for digital images to lower their storage and/or transmission requirements. Source: Variable Rate Dee...	Image Compression is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Point Cloud Registration	C	Point Cloud Registration is a fundamental problem in 3D computer vision and photogrammetry. Given several sets of points in different coordinate systems, the aim of registration is to find the tra...	Point Cloud Registration is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Real-Time Semantic Segmentation	D	N/A	Real-Time Semantic Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Fine-Grained Image Recognition	D	N/A	Fine-Grained Image Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Object Detection In Aerial Images	A	N/A	Object Detection In Aerial Images requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Activity Detection	A	Detecting activities in extended videos.	Activity Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Fraud Detection	X	Fraud Detection is a vital topic that applies to many industries including the financial sectors, banking, government agencies, insurance, and law enforcement, and more. Fraud endeavors have detec...	Fraud Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Video Segmentation	A	N/A	Video Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Surface Normals Estimation	B	Surface normal estimation deals with the task of predicting the surface orientation of the objects present inside a scene. Refer to Designing Deep Networks for Surface Normal Estimation (Wang et al.)...	Surface Normals Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Conversational Response Selection	X	Conversational response selection refers to the task of identifying the most relevant response to a given input sentence from a collection of sentences.	Conversational Response Selection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Motion Forecasting	A	Motion forecasting is the task of predicting the location of a tracked object in the future	Motion Forecasting requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
License Plate Detection	A	License Plate Detection is an image-processing technology used to identify vehicles by their license plates. This technology is used in various security and traffic applications.	License Plate Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Quality Assessment	A	Video Quality Assessment is a computer vision task aiming to mimic video-based human subjective perception. The goal is to produce a mos score, where higher score indicates better perceptual quality. ...	Video Quality Assessment requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Time Series Prediction	X	The goal of Time Series Prediction is to infer the future values of a time series from the past. Source: Orthogonal Echo State Networks and stochastic evalua...	Time Series Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Fact Checking	A	N/A	Fact Checking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Key Information Extraction	A	Key Information Extraction (KIE) is aimed at extracting structured information (e.g. key-value pairs) from form-style documents (e.g. invoices), which makes an important step towards intelligent docum...	Key Information Extraction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Zero-Shot Composed Image Retrieval (ZS-CIR)	A	N/A	Zero-Shot Composed Image Retrieval (ZS-CIR) requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Outlier Detection	D	Outlier Detection is a task of identifying a subset of a given data set which are considered anomalous in that they are unusual from other instances. It is one of the core data mining tasks and is...	Outlier Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...

Archived task	Level	Archived description	Manual interface judgment
Scene Graph Generation	A	A scene graph is a structured representation of an image, where nodes in a scene graph correspond to object bounding boxes with their object categories, and edges correspond to their pairwise relation...	Scene Graph Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Depth Completion	B	The Depth Completion task is a subproblem of depth estimation. In the sparse-to-dense depth completion problem, one wants to infer the dense depth map of a 3-D scene given an RGB image and its co...	Depth Completion is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Graph Matching	A	Graph Matching is the problem of finding correspondences between two sets of vertices while preserving complex relational information among them. Since the graph structure has a strong capacity to...	Graph Matching requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Multimodal Emotion Recognition	C	This is a leaderboard for multimodal emotion recognition on the IEMOCAP dataset. The modality abbreviations are A: Acoustic T: Text V: Visual Please include the modality in the bracket after the...	Multimodal Emotion Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Head Pose Estimation	B	N/A	Head Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Unsupervised Object Segmentation	D	Image credit: ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation	Unsupervised Object Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Referring Expression Segmentation	A	The task aims at labeling the pixels of an image or video that represent an object instance referred by a linguistic expression. In particular, the referring expression (RE) must allow the identificat...	Referring Expression Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
3D Face Reconstruction	A	3D Face Reconstruction is a computer vision task that involves creating a 3D model of a human face from a 2D image or a set of images. The goal of 3D face reconstruction is to reconstruct a digita...	3D Face Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Action Anticipation	A	Next action anticipation is defined as observing 1, ... , T frames and predicting the action that happens after a gap of T_a seconds. It is important to note that a new action starts after T_a seconds...	Action Anticipation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Bias Detection	A	Bias detection is the task of detecting and measuring racism, sexism and otherwise discriminatory behavior in a model.	Bias Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Medical Visual Question Answering	A	N/A	Medical Visual Question Answering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Image Manipulation	B	Image Manipulation is the process of altering or transforming an existing image to achieve a desired effect or to modify its content. This can involve various techniques and tools to enhance, modify, ...	Image Manipulation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Dimensionality Reduction	A	Dimensionality reduction is the task of reducing the dimensionality of a dataset. (Image credit: openTSNE)¡/sp...	Dimensionality Reduction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Event-based vision	C	An event camera, also known as a neuromorphic camera, silicon retina or dynamic vision sensor, is an imaging sensor that responds to local changes in brightness. Event cameras do not capture images us...	Event-based vision is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Material Recognition	B	Material recognition focuses on identifying classes, types, states, and properties of materials.	Material Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Visual Grounding	A	Visual Grounding (VG) aims to locate the most relevant object or region in an image, based on a natural language query. The query can be a phrase, a sentence, or even a multi-round dialogue. There are...	Visual Grounding requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Multi-Label Image Classification	B	The Multi-Label Image Classification focuses on predicting labels for images in a multi-class classification problem where each image may belong to more than one class.	Multi-Label Image Classification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Human Part Segmentation	B	N/A	Human Part Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Unsupervised Object Detection	A	N/A	Unsupervised Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Table Detection	A	Image credit:Table Detection in the Wild: A Novel Diverse Table Detection Dataset and Method	Table Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Passage Retrieval	X	N/A	Passage Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Out of Distribution (OOD) Detection	D	N/A	Out of Distribution (OOD) Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Video Grounding	A	Video grounding is the task of linking spoken language descriptions to specific video segments. In video grounding, the model is given a video and a natural language description, such as a sentenc...	Video Grounding requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Video Recognition	A	N/A	Video Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Online Multi-Object Tracking	A	N/A	Online Multi-Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Scene Segmentation	B	Scene segmentation is the task of splitting a scene into its various object components. Image adapted from Temporally coherent 4D reconstruction of complex dynamic scenes	Scene Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Face Swapping	B	N/A	Face Swapping is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Facial Attribute Classification	B	N/A	Facial Attribute Classification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Answer Selection	X	N/A	Answer Selection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
LIDAR Semantic Segmentation	C	N/A	LIDAR Semantic Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Visual Dialog	A	Visual Dialog requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a follow-up que...	Visual Dialog requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Motion Estimation	A	Motion Estimation is used to determine the block-wise or pixel-wise motion vectors between two frames. Source: MEMC-Net: Motion Estimation and Motion Compens...	Motion Estimation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Acoustic Scene Classification	X	The goal of acoustic scene classification is to classify a test recording into one of the provided predefined classes that characterizes the environment in which it was recorded. Source: DCASE 201...	Acoustic Scene Classification is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Object Detection In Indoor Scenes	A	N/A	Object Detection In Indoor Scenes requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Semantic SLAM	A	N/A	Semantic SLAM requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
10-shot image generation	D	Generate 10 image of famous arts	10-shot image generation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Multiview Detection	A	N/A	Multiview Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Remote Sensing Image Classification	C	N/A	Remote Sensing Image Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
3D Depth Estimation	B	Image: monodepth2	3D Depth Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Font Recognition	B	Font recognition (also called *visual font recognition* or *optical font recognition*) is the task of identifying the font family or families used in images containing text. Understanding which fonts ...	Font Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Time Series Anomaly Detection	X	N/A	Time Series Anomaly Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Within-Session Motor Imagery (left hand vs. right hand)	C	N/A	Within-Session Motor Imagery (left hand vs. right hand) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing,...
Color Image Denoising	B	N/A	Color Image Denoising is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Video Description	A	N/A	Video Description requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Unsupervised Semantic Segmentation	D	Models that learn to segment each image (i.e. assign a class to every pixel) without seeing the ground truth labels. (Image credit: SegSort: Segmentation by...	Unsupervised Semantic Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Conditional Image Generation	B	Conditional image generation is the task of generating new images from a dataset conditional on their class. (Image credit: PixelCNN++	Conditional Image Generation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...

Archived task	Level	Archived description	Manual interface judgment
Interactive Segmentation	A	N/A	Interactive Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Real-Time Object Detection	A	N/A	Real-Time Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Robust Object Detection	A	N/A	Robust Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Edge Detection	B	Edge Detection is a fundamental image processing technique which involves computing an image gradient to quantify the magnitude and direction of edges in an image. Image gradients are used in vari...	Edge Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Scene Generation	B	make to t shirt an Ad with a little bit of action	Scene Generation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Vehicle Re-Identification	A	Vehicle re-identification is the task of identifying the same vehicle across multiple cameras. (Image credit: A Two-Stream Siamese Neural Network for Vehicl...	Vehicle Re-Identification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Small Data Image Classification	D	N/A	Small Data Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Fine-Grained Visual Recognition	D	N/A	Fine-Grained Visual Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Gait Recognition	A	(Image credit: GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition)	Gait Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
6D Pose Estimation using RGBD	B	N/A	6D Pose Estimation using RGBD is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
3D Shape Reconstruction	A	Image credit: GSNet: Joint Vehicle Pose and Shape Reconstruction with Geometrical and Scene-aware Supervision , ECCV'20	3D Shape Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Pedestrian Attribute Recognition	D	Pedestrian attribution recognition is the task of recognizing pedestrian features - such as whether they are talking on a phone, whether they have a backpack, and so on. ¡span style="color:grey; op...	Pedestrian Attribute Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Action Quality Assessment	A	Assessing/analyzing/quantifying how well an action was performed.	Action Quality Assessment requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Reconstruction	A	Source: Deep-SloMo	Video Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Trajectory Planning	A	Trajectory planning for industrial robots consists of moving the tool center point from point A to point B while avoiding body collisions over time. Trajectory planning is sometimes referred to as mo...	Trajectory Planning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Defect Detection	D	N/A	Defect Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
Vision-Language Navigation	A	Vision-language navigation (VLN) is the task of navigating an embodied agent to carry out natural language instructions inside real 3D environments. (Image c...	Vision-Language Navigation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Dense Video Captioning	A	Most natural videos contain numerous events. For example, in a video of a "man playing a piano", the video might also contain "another man dancing" or "a crowd clapping". The task of dense video capti...	Dense Video Captioning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Saliency Prediction	B	A saliency map is a model that predicts eye fixations on a visual scene. Saliency prediction is informed by the human visual attention mechanism and predicts the possibility of the human eyes to stay ...	Saliency Prediction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Blind Super-Resolution	D	N/A	Blind Super-Resolution adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Intrusion Detection	X	Intrusion Detection is the process of dynamically monitoring events occurring in a computer system or network, analyzing them for signs of possible incidents and often interdicting the unauthorize...	Intrusion Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Medical Report Generation	A	Medical report generation (MRG) is a task which focus on training AI to automatically generate professional report according the input image data. This can help clinicians make faster and more accurat...	Medical Report Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Change Point Detection	X	Change Point Detection is concerned with the accurate detection of abrupt and significant changes in the behavior of a time series. Change point detection is the task of finding changes in the ...	Change Point Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Surface Reconstruction	A	N/A	Surface Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Food Recognition	D	N/A	Food Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
Camera Localization	A	N/A	Camera Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
object-detection	A	N/A	object-detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Sound Event Localization and Detection	X	Given multichannel audio input, a sound event detection and localization (SELD) system outputs a temporal activation track for each of the target sound classes, along with one or more corresponding sp...	Sound Event Localization and Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
3D Medical Imaging Segmentation	C	3D medical imaging segmentation is the task of segmenting medical objects of interest from 3D medical imaging. (Image credit: Elastic Boundary Projection fo...	3D Medical Imaging Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Within-Session Motor Imagery (right hand vs. feet)	C	N/A	Within-Session Motor Imagery (right hand vs. feet) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing,...
Zero-Shot Transfer Image Classification	D	N/A	Zero-Shot Transfer Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
3D Object Reconstruction	A	Image: Choy et al	3D Object Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image-to-Text Retrieval	A	Image-text retrieval is the process of retrieving relevant images based on textual descriptions or finding corresponding textual descriptions for a given image. This task is interdisciplinary, com...	Image-to-Text Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Few-Shot Object Detection	A	Few-Shot Object Detection is a computer vision task that involves detecting objects in images with limited training data. The goal is to train a model on a few examples of each object class and th...	Few-Shot Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Layout-to-Image Generation	A	N/A	Layout-to-Image Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Scene Flow Estimation	A	Optical flow is a two-dimensional motion field in the image plane. It is the projection of the three-dimensional motion of the world. If the world is completely non-rigid, the motions of the points in...	Scene Flow Estimation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
No-Reference Image Quality Assessment	B	An Image Quality Assessment approach where no reference image information is available to the model. Sometimes referred to as Blind Image Quality Assessment (BIQA).	No-Reference Image Quality Assessment is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Video-Based Person Re-Identification	A	N/A	Video-Based Person Re-Identification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Semantic correspondence	B	The task of semantic correspondence aims to establish reliable visual correspondence between different instances of the same object category.	Semantic correspondence is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Visual Commonsense Reasoning	A	N/A	Visual Commonsense Reasoning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Seizure Detection	X	Seizure Detection is a binary supervised classification problem with the aim of classifying between seizure and non-seizure states of a patient. Source: ResOT: ...	Seizure Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Image/Document Clustering	A	N/A	Image/Document Clustering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Anomaly Detection In Surveillance Videos	A	N/A	Anomaly Detection In Surveillance Videos requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Video Instance Segmentation	A	The goal of video instance segmentation is simultaneous detection, segmentation and tracking of instances in videos. In words, it is the first time that the image instance segmentation problem is exte...	Video Instance Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Click-Through Rate Prediction	X	Click-through rate prediction is the task of predicting the likelihood that something on a website (such as an advertisement) will be clicked. (Image credit:...	Click-Through Rate Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Unsupervised Video Object Segmentation	A	N/A	Unsupervised Video Object Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Video Semantic Segmentation	A	The goal of video semantic segmentation is to assign a predefined class to each pixel in all frames of a video. This requires the model not only to predict accurate segmentation masks but also to ensu...	Video Semantic Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Weakly-Supervised Semantic Segmentation	D	N/A	Weakly-Supervised Semantic Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
3D Absolute Human Pose Estimation	B	This task aims to solve absolute (camera-centric not root-relative) 3D human pose estimation. (Image credit: RootNet	3D Absolute Human Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Human action generation	A	Yan et al. (2019) CSGN: "When the dancer is stepping, jumping and spinning on the stage, attentions of all audiences are attracted by the streamof the fluent and graceful movements. Building a mo...	Human action generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Retinal Vessel Segmentation	C	Retinal vessel segmentation is the task of segmenting vessels in retina imagery. (Image credit: LadderNet)¡/s...	Retinal Vessel Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Dynamic Link Prediction	X	N/A	Dynamic Link Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Human Interaction Recognition	A	Human Interaction Recognition (HIR) is a field of study that involves the development of computer algorithms to detect and recognize human interactions in videos, images, or other multimedia content. ...	Human Interaction Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Pose Prediction	A	N/A	Pose Prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
geo-localization	A	N/A	geo-localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Image Matting	B	Image Matting is the process of accurately estimating the foreground object in images and videos. It is a very important technique in image and video editing applications, particularly in film pro...	Image Matting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Colorectal Polyps Characterization	C	N/A	Colorectal Polyps Characterization is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Zero-shot Image Retrieval	A	N/A	Zero-shot Image Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Surgical phase recognition	A	The first 40 videos are used for training, the last 40 videos are used for testing.	Surgical phase recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Satellite Image Classification	C	N/A	Satellite Image Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Single Image Dehazing	B	N/A	Single Image Dehazing is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
3D Object Recognition	C	N/A	3D Object Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Intrinsic Image Decomposition	B	Intrinsic Image Decomposition is the process of separating an image into its formation components such as reflectance (albedo) and shading (illumination). Reflectance is the color of the object, i...	Intrinsic Image Decomposition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Keyphrase Extraction	X	A classic task to extract salient phrases that best summarize a document, which essentially has two stages: candidate generation and keyphrase ranking.	Keyphrase Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Monocular Visual Odometry	A	N/A	Monocular Visual Odometry requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Ad-hoc video search	A	The Ad-hoc search task ended a 3 year cycle from 2016-2018 with a goal to model the end user search use-case, who is searching (using textual sentence queries) for segments of video containing persons...	Ad-hoc video search requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Medical Image Registration	C	Image registration, also known as image fusion or image matching, is the process of aligning two or more images based on image appearances. Medical Image Registration seeks to find an optimal spat...	Medical Image Registration is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Skin Lesion Classification	C	N/A	Skin Lesion Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...

Archived task	Level	Archived description	Manual interface judgment
Brain Tumor Segmentation	C	N/A	Brain Tumor Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Entity Alignment	A	Entity Alignment is the task of finding entities in two knowledge bases that refer to the same real-world object. It plays a vital role in automatically integrating multiple knowledge bases. No...	Entity Alignment requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Gender Prediction	B	N/A	Gender Prediction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image Forensics	B	N/A	Image Forensics is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Colorization	B	Colorization is the process of adding plausible color information to monochrome photographs or videos. Colorization is a highly undetermined problem, requiring mapping a real-valued luminance imag...	Colorization is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Image Manipulation Localization	B	The task of segmenting parts of images or image parts that have been tampered with or manipulated (sometimes also referred to as doctored). This typically encompasses image splicing, copy-move, or ima...	Image Manipulation Localization is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Product Recommendation	X	N/A	Product Recommendation is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Breast Cancer Detection	C	N/A	Breast Cancer Detection is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Dialogue Act Classification	X	N/A	Dialogue Act Classification is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Handwritten Digit Recognition	B	N/A	Handwritten Digit Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Zero-shot Text-to-Image Retrieval	A	N/A	Zero-shot Text-to-Image Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Multi-Person Pose Estimation	B	Multi-person pose estimation is the task of estimating the pose of multiple people in one frame. (Image credit: Human Pose Estimation with TensorFlow	Multi-Person Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Unsupervised Semantic Segmentation with Language-image Pre-training	A	N/A	Unsupervised Semantic Segmentation with Language-image Pre-training requires language, ranked retrieval, tracks/instances, navigation/trajectory,...
Open Vocabulary Object Detection	A	Open-vocabulary detection (OVD) aims to generalize beyond the limited number of base classes labeled during the training phase. The goal is to detect novel classes defined by an unbounded (open) voca...	Open Vocabulary Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Audio-Visual Speech Recognition	A	Audio-visual speech recognition is the task of transcribing a paired audio and visual stream into text.	Audio-Visual Speech Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Monocular 3D Object Detection	A	N/A	Monocular 3D Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Stereo Disparity Estimation	B	N/A	Stereo Disparity Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Medical Code Prediction	X	Context: Prediction of medical codes from clinical notes is both a practical and essential need for every healthcare delivery organization within current medical systems. Automating annotation will sa...	Medical Code Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Zero-Shot Video Retrieval	A	Zero-shot video retrieval is the task of retrieving relevant videos based on a query (usually in text form) without any prior training on specific examples of those videos. Unlike traditional retrieva...	Zero-Shot Video Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Text to Video Retrieval	A	She's gone I can't find her anywhere I'm looking everywhere for her Everywhere is dark	Text to Video Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
6D Pose Estimation using RGB	B	N/A	6D Pose Estimation using RGB is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Human Pose Forecasting	A	N/A	Human Pose Forecasting requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Face Identification	B	N/A	Face Identification is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Traffic Sign Recognition	D	N/A	Traffic Sign Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Motion Segmentation	A	Motion Segmentation is an essential task in many applications in Computer Vision and Robotics, such as surveillance, action recognition and scene understanding. The classic way to state the proble...	Motion Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Facial Expression Recognition	B	N/A	Facial Expression Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
3D Shape Modeling	A	Image: Gkioxari et al	3D Shape Modeling requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Video Denoising	A	N/A	Video Denoising requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Camouflaged Object Segmentation	D	Camouflaged object segmentation (COS) or Camouflaged object detection (COD), which was originally promoted by T.-N. Le et al....	Camouflaged Object Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Electroencephalogram (EEG)	C	Electroencephalogram (EEG) is a method of recording brain activity using electrophysiological indexes. When the brain is active, a large number of postsynaptic potentials generated synchronously b...	Electroencephalogram (EEG) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Keyword Extraction	X	Keyword extraction is tasked with the automatic identification of terms that best describe the subject of a document (Source: Wikipedia).	Keyword Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
3D Object Classification	C	3D Object Classification is the task of predicting the class of a 3D object point cloud. It is a voxel level prediction where each voxel is classified into a category. The popular benchmark for this ...	3D Object Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Camera Relocalization	A	”Camera relocalization, or image-based localization is a fundamental problem in robotics and computer vision. It refers to the process of determining camera pose from the visual scene representation a...	Camera Relocalization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Lesion Classification	C	N/A	Lesion Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Text Segmentation	X	N/A	Text Segmentation is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Natural Language Visual Grounding	A	N/A	Natural Language Visual Grounding requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Text Matching	X	Matching a target text to a source text based on their meaning.	Text Matching is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Traffic Sign Detection	A	N/A	Traffic Sign Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Gesture Generation	A	N/A	Gesture Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Hyperspectral Image Segmentation	C	N/A	Hyperspectral Image Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Person Recognition	B	N/A	Person Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
3D Human Shape Estimation	A	N/A	3D Human Shape Estimation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Generalizable Novel View Synthesis	A	N/A	Generalizable Novel View Synthesis requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Neural Rendering	A	Given a representation of a 3D scene of some kind (point cloud, mesh, voxels, etc.), the task is to create an algorithm that can produce photorealistic renderings of this scene from an arbitrary viewp...	Neural Rendering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Gender Bias Detection	A	N/A	Gender Bias Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Table Recognition	A	Table recognition refers to the process of automatically identifying and extracting tabular structures from unstructured data sources such as text documents, images, or scanned documents. The goal of ...	Table Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Fault Detection	X	N/A	Fault Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Vulnerability Detection	X	Vulnerability detection plays a crucial role in safeguarding against these threats by identifying weaknesses and potential entry points that malicious actors could exploit. Through advanced scanning t...	Vulnerability Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Infrared And Visible Image Fusion	C	Image fusion with paired infrared and visible images	Infrared And Visible Image Fusion is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Term Extraction	X	Term Extraction, or Automated Term Extraction (ATE), is about extraction domain-specific terms from natural language text. For example, the sentence “We meta-analyzed mortality using random-effect m...	Term Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.

Archived task	Level	Archived description	Manual interface judgment
MRI segmentation	C	N/A	MRI segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Video Restoration	A	N/A	Video Restoration requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
2D Cyclist Detection	A	N/A	2D Cyclist Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Text-based Image Editing	A	N/A	Text-based Image Editing requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image to Video Generation	A	Image to Video Generation refers to the task of generating a sequence of video frames based on a single still image or a set of still images. The goal is to produce a video that is coherent and co...	Image to Video Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Unsupervised Image Classification	D	N/A	Unsupervised Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Structured Prediction	A	Structured Prediction is an area of machine learning focusing on representations of spaces with combinatorial structure, and algorithms for inference and parameter estimation over these structures...	Structured Prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image Deblurring	B	No archived description.	Image Deblurring is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Lightweight Face Recognition	D	N/A	Lightweight Face Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
3D Part Segmentation	C	N/A	3D Part Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Real-time Instance Segmentation	A	N/A	Real-time Instance Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Zero-Shot Object Detection	A	N/A	Zero-Shot Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Open World Object Detection	A	N/A	Open World Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Talking Face Generation	A	Talking face generation aims to synthesize a sequence of face images that correspond to given speech semantics (Image credit: Talking Face Generation by A...	Talking Face Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Stereo Depth Estimation	B	N/A	Stereo Depth Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Egocentric Pose Estimation	B	N/A	Egocentric Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Depth Prediction	B	N/A	Depth Prediction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Zero-Shot Action Recognition	A	N/A	Zero-Shot Action Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Few Shot Action Recognition	A	N/A	Few Shot Action Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Scene Graph Detection	A	N/A	Scene Graph Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Open Vocabulary Semantic Segmentation	A	Open-vocabulary semantic segmentation models aim to accurately assign a semantic label to each pixel in an image from a set of arbitrary open-vocabulary texts.	Open Vocabulary Semantic Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Face Sketch Synthesis	B	N/A	Face Sketch Synthesis is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Medical Named Entity Recognition	X	N/A	Medical Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Moment Retrieval	A	Moment retrieval can de defined as the task of ”localizing moments in a video given a user query”. Description from: QVHIGHLIGHTS: Detecting Moments and Highlights in Videos via Natural Language Q...	Moment Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Grammatical Error Detection	X	N/A	Grammatical Error Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Facial Emotion Recognition	B	N/A	Facial Emotion Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Multimodal Activity Recognition	A	N/A	Multimodal Activity Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Aesthetics Quality Assessment	B	N/A	Aesthetics Quality Assessment is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Spatio-Temporal Action Localization	A	N/A	Spatio-Temporal Action Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
3D Facial Landmark Localization	B	N/A	3D Facial Landmark Localization is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...

Archived task	Level	Archived description	Manual interface judgment
Chinese Named Entity Recognition	X	N/A	Chinese Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Weakly-Supervised Named Entity Recognition	X	N/A	Weakly-Supervised Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Body Detection	A	N/A	Body Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Speaker Recognition	C	Speaker Recognition is the process of identifying or confirming the identity of a person given his speech segments. Source: Margin Matters: Towards More Discrim...	Speaker Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Medical Image Generation	C	Medical image generation is the task of synthesising new medical images. (Image credit: Towards Adversarial Retinal Image Synthesis	Medical Image Generation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
3D Multi-Object Tracking	A	Image: Weng et al	3D Multi-Object Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Sleep Stage Detection	X	Human Sleep Staging into W-N1-N2-N3-REM classes from multiple or single polysomnography signals	Sleep Stage Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Unified Image Restoration	B	Using a single model to restore inputs with different degradation types.	Unified Image Restoration is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Image-Based Localization	A	N/A	Image-Based Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
3D Point Cloud Classification	C	N/A	3D Point Cloud Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Visual Speech Recognition	A	N/A	Visual Speech Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Line Segment Detection	A	N/A	Line Segment Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
3D Face Animation	A	Image: Cudeiro et al	3D Face Animation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Dictionary Learning	D	Dictionary Learning is an important problem in multiple areas, ranging from computational neuroscience, machine learning, to computer vision and image processing. The general goal is to find a goo...	Dictionary Learning adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant...
Robust Face Recognition	D	N/A	Robust Face Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Stereo Matching Hand	B	N/A	Stereo Matching Hand is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Zero-Shot Image Classification	D	Zero-shot image classification is a technique in computer vision where a model can classify images into categories that were not present during training. This is achieved by leveraging semantic inform...	Zero-Shot Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Scene Change Detection	A	Scene change detection (SCD) refers to the task of localizing changes and identifying change-categories given two scenes. A scene can be either an RGB (+D) image or a 3D reconstruction (point cloud)....	Scene Change Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Enhancement	A	N/A	Video Enhancement requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Gaze Prediction	B	N/A	Gaze Prediction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Emotion Recognition in Context	A	N/A	Emotion Recognition in Context requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Composed Image Retrieval (CoIR)	A	Composed Image Retrieval (CoIR) is the task involves retrieving images from a large database based on a query composed of multiple elements, such as text, images, and sketches. The goal is to deve...	Composed Image Retrieval (CoIR) requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
CAD Reconstruction	A	N/A	CAD Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Autonomous Navigation	A	Autonomous navigation is the task of autonomously navigating a vehicle or robot to or around a location without human guidance. (Image credit: Approximate L...	Autonomous Navigation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Single Image Deraining	B	N/A	Single Image Deraining is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Abstractive Dialogue Summarization	X	N/A	Abstractive Dialogue Summarization is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
The Semantic Segmentation Of Remote Sensing Imagery	C	N/A	The Semantic Segmentation Of Remote Sensing Imagery is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing,...

Archived task	Level	Archived description	Manual interface judgment
Trajectory Modeling	A	The equivalent of language modeling but for trajectories.	Trajectory Modeling requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Panoptic Segmentation	A	N/A	Video Panoptic Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video-Text Retrieval	A	N/A	Video-Text Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Multiple People Tracking	A	N/A	Multiple People Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Object SLAM	A	N/A	Object SLAM requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured logic;...
Medical Object Detection	A	N/A	Medical Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Attribute Value Extraction	A	Attribute Value Extraction is the task of extracting values for a given set of attributes of interest from free text input. Attribute value extraction is for example applied in the context of e-co...	Attribute Value Extraction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Within-Session Motor Imagery (all classes)	C	N/A	Within-Session Motor Imagery (all classes) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
Automatic Phoneme Recognition	X	N/A	Automatic Phoneme Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Unsupervised Anomaly Detection with Specified Settings – 20% anomaly	D	N/A	Unsupervised Anomaly Detection with Specified Settings – 20% anomaly adds no new modality or output structure and is mainly a domain,...
Unsupervised Anomaly Detection with Specified Settings – 1% anomaly	D	N/A	Unsupervised Anomaly Detection with Specified Settings – 1% anomaly adds no new modality or output structure and is mainly a domain, supervision,...
Unsupervised Anomaly Detection with Specified Settings – 0.1% anomaly	D	N/A	Unsupervised Anomaly Detection with Specified Settings – 0.1% anomaly adds no new modality or output structure and is mainly a domain,...
Unsupervised Anomaly Detection with Specified Settings – 10% anomaly	D	N/A	Unsupervised Anomaly Detection with Specified Settings – 10% anomaly adds no new modality or output structure and is mainly a domain,...
Blind Face Restoration	B	Blind face restoration aims at recovering high-quality faces from the low-quality counterparts suffering from unknown degradation, such as low-resolution, noise, blur, compression artifacts, etc. When...	Blind Face Restoration is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
3D Face Modelling	B	N/A	3D Face Modelling is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Zero-Shot Cross-Modal Retrieval	A	N/A	Zero-Shot Cross-Modal Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
JPEG Artifact Correction	B	N/A	JPEG Artifact Correction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Stereo Image Super-Resolution	B	N/A	Stereo Image Super-Resolution is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Unsupervised Panoptic Segmentation	A	Unsupervised Panoptic Segmentation aims to partition an image into semantically meaningful regions and distinct object instances without training using any manually annotated images. ¡span style=...	Unsupervised Panoptic Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Unsupervised Monocular Depth Estimation	D	N/A	Unsupervised Monocular Depth Estimation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Self-Supervised Action Recognition	A	N/A	Self-Supervised Action Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Visual Relationship Detection	A	Visual relationship detection (VRD) is one newly developed computer vision task aiming to recognize relations or interactions between objects in an image. It is a further learning task after object re...	Visual Relationship Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Sketch-to-Image Translation	B	N/A	Sketch-to-Image Translation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Pose Transfer	A	N/A	Pose Transfer requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Weakly Supervised Action Localization	A	In this task, the training data consists of videos with a list of activities in them without any temporal boundary annotations. However, while testing, given a video, the algorithm should recognize th...	Weakly Supervised Action Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Oriented Object Detection	A	N/A	Oriented Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Aspect Extraction	X	Aspect extraction is the task of identifying and extracting terms relevant for opinion mining and sentiment analysis, for example terms for product attributes or features.	Aspect Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Sports Ball Detection and Tracking	A	N/A	Sports Ball Detection and Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Symmetry Detection	B	N/A	Symmetry Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Interactive Video Object Segmentation	A	N/A	Interactive Video Object Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Monocular 3D Human Pose Estimation	B	N/A	Monocular 3D Human Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
2D Panoptic Segmentation	A	N/A	2D Panoptic Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Zero-shot 3D Point Cloud Classification	C	N/A	Zero-shot 3D Point Cloud Classification is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
Facial Inpainting	B	N/A	Facial Inpainting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Optic Disc Segmentation	C	N/A	Optic Disc Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Aspect Category Detection	X	Aspect category detection (ACD) in sentiment analysis aims to identify the aspect categories mentioned in a sentence.	Aspect Category Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Grasp Contact Prediction	A	N/A	Grasp Contact Prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Natural Language Moment Retrieval	A	N/A	Natural Language Moment Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
3D Multi-Person Pose Estimation	B	This task aims to solve root-relative 3D multi-person pose estimation. No human bounding box and root joint coordinate groundtruth are used in testing time. (...	3D Multi-Person Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Face Presentation Attack Detection	B	N/A	Face Presentation Attack Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Joint Demosaicing and Denoising	B	N/A	Joint Demosaicing and Denoising is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Motion Planning	A	(Image credit: Motion Planning Among Dynamic, Decision-Making Agents with Deep Reinforcement Learning)	Motion Planning requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
3D Semantic Scene Completion	C	This task was introduced in "Semantic Scene Completion from a Single Depth Image" at CVPR 2017 . The target is to infer the dense 3D voxelized semantic scene from an...	3D Semantic Scene Completion is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Zeroshot Video Question Answer	A	N/A	Zeroshot Video Question Answer requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Depression Detection	A	N/A	Depression Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Fine-Grained Visual Categorization	D	N/A	Fine-Grained Visual Categorization adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
RGB-D Salient Object Detection	B	N/A	RGB-D Salient Object Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Extractive Text Summarization	X	Given a document, selecting a subset of the words or sentences which best represents a summary of the document.	Extractive Text Summarization is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Stock Market Prediction	X	N/A	Stock Market Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Cross-Domain Few-Shot Object Detection	A	N/A	Cross-Domain Few-Shot Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Surgical tool detection	A	N/A	Surgical tool detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Hand-Gesture Recognition	A	N/A	Hand-Gesture Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Surface Normal Estimation	B	N/A	Surface Normal Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Face Parsing	B	N/A	Face Parsing is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Face Model	A	N/A	Face Model requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured logic;...
Content-Based Image Retrieval	A	Content-Based Image Retrieval is a well studied problem in computer vision, with retrieval problems generally divided into two groups: category-level retrieval and instance-level retrieval. Given ...	Content-Based Image Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Medical Relation Extraction	X	Biomedical relation extraction is the task of detecting and classifying semantic relationships from biomedical text.	Medical Relation Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Text to Audio Retrieval	X	N/A	Text to Audio Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Video Emotion Recognition	A	N/A	Video Emotion Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
3D Feature Matching	A	N/A	3D Feature Matching requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Low Resource Named Entity Recognition	X	N/A	Low Resource Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Video Compression	A	N/A	Video Compression requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Program Synthesis	X	Program synthesis is the process of automatically generating a program or code snippet that satisfies a given specification or set of requirements. This can include generating code from a formal speci...	Program Synthesis is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Malware Detection	X	N/A	Malware Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Human motion prediction	A	Action prediction is a pre-fact video understanding task, which focuses on future states, in other words, it needs to reason about future states or infer action labels before the end of action executi...	Human motion prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Face Generation	B	Face generation is the task of generating (or interpolating) new faces from an existing dataset. The state-of-the-art results for this task are located in the Image Generation parent. ¡span styl...	Face Generation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
whole slide images	C	N/A	whole slide images is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Logo Recognition	D	N/A	Logo Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
3D Place Recognition	A	N/A	3D Place Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Cell Detection	A	N/A	Cell Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Spatial Relation Recognition	A	N/A	Spatial Relation Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image Relighting	B	N/A	Image Relighting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Trajectory Clustering	A	N/A	Trajectory Clustering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Music Emotion Recognition	X	N/A	Music Emotion Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Medical X-Ray Image Segmentation	C	N/A	Medical X-Ray Image Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
3D Assembly	A	N/A	3D Assembly requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured logic;...
Brain Segmentation	C	N/A	Brain Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Medical Concept Normalization	X	N/A	Medical Concept Normalization is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Road Segmentation	D	Road Segmentation is a pixel wise binary classification in order to extract underlying road network. Various Heuristic and data driven models are proposed. Continuity and robustness still remains one ...	Road Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
Photoplethysmography (PPG)	C	Photoplethysmography (PPG) is a non-invasive light-based method that has been used since the 1930s for monitoring cardiovascular activity. Source: Non-contac...	Photoplethysmography (PPG) is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
3D Anomaly Detection	C	3D-only Anomaly Detection. Structures out of normal distribution are detected from the 3D-only point cloud.	3D Anomaly Detection is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Photoplethysmography (PPG) heart rate estimation	C	N/A	Photoplethysmography (PPG) heart rate estimation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
3D Question Answering (3D-QA)	A	N/A	3D Question Answering (3D-QA) requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Text to 3D	A	Task involves generating 3D objects based on the text prompt provided to the system.	Text to 3D requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured logic;...
Video-Adverb Retrieval	A	The bidirectional video-adverb retrieval task aims at retrieving adverbs that match an action in a video and vice versa.	Video-Adverb Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...

Archived task	Level	Archived description	Manual interface judgment
Image Forgery Detection	B	N/A	Image Forgery Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Cross-modal retrieval with noisy correspondence	A	N/A	Cross-modal retrieval with noisy correspondence requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
3D Generation	A	N/A	3D Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Handwritten Mathmatical Expression Recognition	A	Offline Handwritten mathematical Expression Recognition aims to convert 2D images into a 1D structured sequences(LaTeX or MathML)	Handwritten Mathmatical Expression Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Image-text Retrieval	A	N/A	Image-text Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Multiview Clustering	A	N/A	Multiview Clustering requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Unsupervised Anomaly Detection with Specified Settings – 30% anomaly	D	N/A	Unsupervised Anomaly Detection with Specified Settings – 30% anomaly adds no new modality or output structure and is mainly a domain,...
Image Compressed Sensing	B	N/A	Image Compressed Sensing is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Physical Attribute Prediction	B	N/A	Physical Attribute Prediction is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Multimodal Text and Image Classification	A	Classification with both source Image and Text	Multimodal Text and Image Classification requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Image Outpainting	B	Predicting the visual context of an image beyond its boundary. Image credit: NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis	Image Outpainting is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Vehicle Pose Estimation	B	N/A	Vehicle Pose Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
3D Object Detection From Stereo Images	A	N/A	3D Object Detection From Stereo Images requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Prediction Of Occupancy Grid Maps	C	N/A	Prediction Of Occupancy Grid Maps is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Self-Supervised Action Recognition Linear	A	N/A	Self-Supervised Action Recognition Linear requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Mortality Prediction	X	(Image credit: Early hospital mortality prediction using vital signals)	Mortality Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Unsupervised Image Segmentation	D	N/A	Unsupervised Image Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Patch Matching	B	N/A	Patch Matching is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
Facial Action Unit Detection	B	N/A	Facial Action Unit Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Clinical Concept Extraction	X	Automatic extraction of clinical named entities such as clinical problems, treatments, tests and anatomical parts from clinical notes. (Source)	Clinical Concept Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Temporal Localization	A	N/A	Temporal Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Temporal Action Proposal Generation	A	N/A	Temporal Action Proposal Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Weakly-supervised Temporal Action Localization	A	Temporal Action Localization with weak supervision where only video-level labels are given for training	Weakly-supervised Temporal Action Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Abnormal Event Detection In Video	A	Abnormal Event Detection In Video is a challenging task in computer vision, as the definition of what an abnormal event looks like depends very much on the context. For instance, a car driving by ...	Abnormal Event Detection In Video requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Egocentric Activity Recognition	A	N/A	Egocentric Activity Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Semi-Supervised Instance Segmentation	A	N/A	Semi-Supervised Instance Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Multimodal Unsupervised Image-To-Image Translation	B	N/A	Multimodal Unsupervised Image-To-Image Translation is compatible with existing WSG dense fields or pose representations; only labels, data...
Unsupervised Action Segmentation	A	N/A	Unsupervised Action Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Supervised Video Summarization	A	N/A	Supervised Video Summarization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Video Salient Object Detection	A	N/A	Video Salient Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Image Classification with Label Noise	D	N/A	Image Classification with Label Noise adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Reconstruction	A	N/A	Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Pose Retrieval	A	Retrieval of similar human poses from images or videos	Pose Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...

Archived task	Level	Archived description	Manual interface judgment
Document Image Classification	D	Document image classification is the task of classifying documents based on images of their contents. (Image credit: Real-Time Document Image Classification...	Document Image Classification adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
3D-Aware Image Synthesis	A	N/A	3D-Aware Image Synthesis requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Video Polyp Segmentation	A	N/A	Video Polyp Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Unsupervised Extractive Summarization	X	N/A	Unsupervised Extractive Summarization is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Nested Mention Recognition	X	Nested mention recognition is the task of correctly modeling the nested structure of mentions.	Nested Mention Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Event Argument Extraction	X	N/A	Event Argument Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Image Stitching	A	Image Stitching is a process of composing multiple images with narrow but overlapping fields of view to create a larger image with a wider field of view. Sour...	Image Stitching requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
3D Open-Vocabulary Instance Segmentation	A	Open-vocabulary 3D instance segmentation is a computer vision task that involves identifying and delineating individual objects or instances within a three-dimensional (3D) scene without prior knowled...	3D Open-Vocabulary Instance Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Cross-Domain Named Entity Recognition	X	N/A	Cross-Domain Named Entity Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Cross-View Image-to-Image Translation	A	N/A	Cross-View Image-to-Image Translation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Shadow Detection	B	N/A	Shadow Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Weakly supervised Semantic Segmentation	D	N/A	Weakly supervised Semantic Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
3D Semantic Scene Completion from a single RGB image	C	N/A	3D Semantic Scene Completion from a single RGB image is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing,...
Few-Shot Semantic Segmentation	D	Few-shot semantic segmentation (FSS) learns to segment target objects in query image given few pixel-wise annotated support image.	Few-Shot Semantic Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Artery/Veins Retinal Vessel Segmentation	C	N/A	Artery/Veins Retinal Vessel Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
Drone-view target localization	A	N/A	Drone-view target localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Weakly Supervised Defect Detection	D	N/A	Weakly Supervised Defect Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or...
Nuclear Segmentation	C	N/A	Nuclear Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Action Localization	A	Action Localization is finding the spatial and temporal co ordinates for an action in a video. An action localization model will identify which frame an action start and ends in video and return the x...	Action Localization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Sketch-Based Image Retrieval	A	N/A	Sketch-Based Image Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Visual Storytelling	A	(Image credit: No Metrics Are Perfect)	Visual Storytelling requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Building change detection for remote sensing images	A	N/A	Building change detection for remote sensing images requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Photo geolocation estimation	A	N/A	Photo geolocation estimation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Action Understanding	A	N/A	Action Understanding requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Multiview Learning	D	N/A	Multiview Learning adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
3D Human Pose Tracking	A	N/A	3D Human Pose Tracking requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
MRI Reconstruction	C	In its most basic form, MRI reconstruction consists in retrieving a complex-valued image from its under-sampled Fourier coefficients. Besides, it can be addressed as a encoder-decoder task, in which...	MRI Reconstruction is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...

Archived task	Level	Archived description	Manual interface judgment
Stock Prediction	X	N/A	Stock Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Action Triplet Recognition	A	N/A	Action Triplet Recognition requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Event Coreference Resolution	X	N/A	Event Coreference Resolution is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Change detection for remote sensing images	A	N/A	Change detection for remote sensing images requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or...
Face Reconstruction	A	Face reconstruction is the task of recovering the facial geometry of a face from an image. (Image credit: Microsoft Deep3DFaceReconstruction	Face Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Multi-future Trajectory Prediction	A	N/A	Multi-future Trajectory Prediction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Distant Speech Recognition	X	N/A	Distant Speech Recognition is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Video Alignment	A	N/A	Video Alignment requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
EEG Emotion Recognition	C	N/A	EEG Emotion Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Instrument Recognition	C	N/A	Instrument Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Word Alignment	X	Word Alignment is the task of finding the correspondence between source and target words in a pair of sentences that are translations of each other. Source: ...	Word Alignment is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Scientific Results Extraction	A	N/A	Scientific Results Extraction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Iris Recognition	C	N/A	Iris Recognition is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Supervised Anomaly Detection	D	N/A	Supervised Anomaly Detection adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application...
Scene Parsing	B	Scene parsing is to segment and parse an image into different image regions associated with semantic categories, such as sky, road, person, and bed. MIT Description	Scene Parsing is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
3D Object Retrieval	A	N/A	3D Object Retrieval requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Temporal View Synthesis	A	N/A	Temporal View Synthesis requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Camera shot boundary detection	A	N/A	Camera shot boundary detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Indoor Monocular Depth Estimation	B	N/A	Indoor Monocular Depth Estimation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small...
Audio to Text Retrieval	X	N/A	Audio to Text Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Multi-modal Entity Alignment	A	N/A	Multi-modal Entity Alignment requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Online Beat Tracking	X	N/A	Online Beat Tracking is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Online Downbeat Tracking	X	N/A	Online Downbeat Tracking is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Spelling Correction	X	Spelling correction is the task of detecting and correcting spelling mistakes.	Spelling Correction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Zero-shot Named Entity Recognition (NER)	X	N/A	Zero-shot Named Entity Recognition (NER) is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Image Cropping	B	Image Cropping is a common photo manipulation process, which improves the overall composition by removing unwanted regions. Image Cropping is widely used in photographic, film processing, graphic ...	Image Cropping is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side adjustment...
graph construction	A	N/A	graph construction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Document-level Relation Extraction	X	Document-level RE aim to identify the relations of various entity pairs expressed across multiple sentences.	Document-level Relation Extraction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
3D Shape Representation	B	Image: MeshNet	3D Shape Representation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Hand Segmentation	B	N/A	Hand Segmentation is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Landmark Recognition	D	N/A	Landmark Recognition adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant...

Archived task	Level	Archived description	Manual interface judgment
Abuse Detection	X	Abuse detection is the task of identifying abusive behaviors, such as hate speech, offensive language, sexism and racism, in utterances from social media platforms.	Abuse Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Multispectral Object Detection	A	Only using RGB cameras for automatic outdoor scene analysis is challenging when, for example, facing insufficient illumination or adverse weather. To improve the recognition reliability, multispectral...	Multispectral Object Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Medical Image Enhancement	C	Aims to improve the perceptual quality of low-quality medical images	Medical Image Enhancement is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Action Parsing	A	Action parsing is the task of, given a video or still image, assigning each frame or image a label describing the action in that frame or image.	Action Parsing requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Humor Detection	A	Humor detection is the task of identifying comical or amusing elements.	Humor Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Thermal Image Segmentation	C	N/A	Thermal Image Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral,...
Sketch Recognition	B	N/A	Sketch Recognition is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Cross-View Geo-Localisation	A	N/A	Cross-View Geo-Localisation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
3D Scene Reconstruction	A	Creating 3D scene either using conventional SFM pipelines or latest deep learning approaches.	3D Scene Reconstruction requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Image-based Automatic Meter Reading	A	N/A	Image-based Automatic Meter Reading requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Cross-Lingual Abstractive Summarization	X	N/A	Cross-Lingual Abstractive Summarization is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Extracting Buildings In Remote Sensing Images	C	N/A	Extracting Buildings In Remote Sensing Images is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event,...
Heart Segmentation	C	N/A	Heart Segmentation is interface-compatible, but its medical, scientific, point-cloud, LiDAR, remote-sensing, event, hyperspectral, thermal,...
Action Generation	A	N/A	Action Generation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Malware Family Detection	X	N/A	Malware Family Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Negation Detection	X	Negation detection is the task of identifying negation cues in text.	Negation Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
hand-object pose	B	N/A	hand-object pose is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Biomedical Information Retrieval	X	N/A	Biomedical Information Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Bird Audio Detection	X	N/A	Bird Audio Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Entity Retrieval	X	N/A	Entity Retrieval is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Rumour Detection	X	Rumor detection is the task of identifying rumors, i.e. statements whose veracity is not quickly or ever confirmed, in utterances on social media platforms.	Rumour Detection is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Semantic Image-Text Similarity	A	N/A	Semantic Image-Text Similarity requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Stock Trend Prediction	X	N/A	Stock Trend Prediction is nonvisual in this archive entry, with no intrinsic visual input, output, or visual auxiliary role.
Crack Segmentation	D	N/A	Crack Segmentation adds no new modality or output structure and is mainly a domain, supervision, robustness, deployment, or application variant of...
Hand Detection	A	As an important subject in the field of computer vision, hand detection plays an important role in many tasks such as human-computer interaction, automatic driving, virtual reality and so on.	Hand Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse structured...
Blood Cell Detection	A	N/A	Blood Cell Detection requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
Referring Video Object Segmentation	A	N/A	Referring Video Object Segmentation requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other...
Fake Image Detection	B	(Image credit: FaceForensics++)	Fake Image Detection is compatible with existing WSG dense fields or pose representations; only labels, data adapters, or a small output-side...
Audio-Visual Synchronization	A	N/A	Audio-Visual Synchronization requires language, ranked retrieval, tracks/instances, navigation/trajectory, a mesh/graph, or temporal/other sparse...
audio-visual learning	C	N/A	audio-visual learning remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Real-Time Multi-Object Tracking	A	N/A	Real-Time Multi-Object Tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...

Archived task	Level	Archived description	Manual interface judgment
Intent Recognition	A	N/A	Intent Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Temporal Forgery Localization	A	N/A	Temporal Forgery Localization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Discourse Segmentation	X	N/A	Discourse Segmentation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Temporal Relation Extraction	X	Temporal relation extraction systems aim to identify and classify the temporal relation between a pair of entities provided in a text. For instance, in the sentence "Bob sent a message to Alice while ...	Temporal Relation Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
UNET Segmentation	D	N/A	UNET Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
3D Lane Detection	C	N/A	3D Lane Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Camera Calibration	C	Camera calibration involves estimating camera parameters(including camera intrinsics and extrinsics) to infer geometric features from captured sequences, which is crucial for computer vision and robot...	Camera Calibration remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Robust Speech Recognition	X	N/A	Robust Speech Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Optical Charater Recogntion	B	N/A	Optical Charater Recogntion maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Image Defocus Deblurring	B	N/A	Image Defocus Deblurring maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
3D Facial Expression Recognition	C	N/A	3D Facial Expression Recognition remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Cell Entity Annotation	X	N/A	Cell Entity Annotation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Person Retrieval	A	N/A	Person Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Lesion Detection	C	N/A	Lesion Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Dynamic Facial Expression Recognition	A	N/A	Dynamic Facial Expression Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Point Tracking	A	Point Tracking, often referred to as Tracking any Point (TAP) involves acquiring, focusing on, and continuously tracking specific target point/points across video frames. The system identifies the tar...	Point Tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Financial Relation Extraction	X	N/A	Financial Relation Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Social Navigation	A	N/A	Social Navigation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Synthetic Speech Detection	X	N/A	Synthetic Speech Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
medical image detection	C	N/A	medical image detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Spectral Reconstruction	C	N/A	Spectral Reconstruction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Event-based Object Segmentation	A	N/A	Event-based object segmentation consumes an asynchronous temporal event stream; temporal structure takes precedence over merely adding the event...
Deformable Object Manipulation	A	N/A	Deformable Object Manipulation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
2D Tiny Object Detection	D	N/A	2D Tiny Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Hyperspectral image analysis	C	N/A	Hyperspectral image analysis remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Video Editing	A	N/A	Video Editing contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Sequential Image Classification	D	N/A	Sequential Image Classification is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Image Colorization	B	N/A	Image Colorization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Image Attribution	B	Image attribution algorithms aim to identify important regions that are highly relevant to model decisions.	Image Attribution maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Model Compression	X	Model Compression is an actively pursued area of research over the last few years with the goal of deploying state-of-the-art deep networks in low-power and resource limited devices without signif...	Model Compression has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Weakly-Supervised Object Localization	D	N/A	Weakly-Supervised Object Localization is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Self-Supervised Image Classification	D	N/A	Self-Supervised Image Classification is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.

Archived task	Level	Archived description	Manual interface judgment
Synthetic Face Recognition	D	N/A	Synthetic Face Recognition is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Face Quality Assessement	D	N/A	Face Quality Assessement is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Face Anonymization	B	N/A	Face Anonymization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Semi Supervised Learning for Image Captioning	A	N/A	Semi Supervised Learning for Image Captioning contains visual evidence but needs structured language, temporal composition, retrieval, or control...
mage-to-Text Retrieval	A	N/A	mage-to-Text Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Single-object discovery	B	N/A	Single-object discovery maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Multi-object discovery	B	N/A	Multi-object discovery maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Object Proposal Generation	B	N/A	Object Proposal Generation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Visual Keyword Spotting	A	N/A	Visual Keyword Spotting contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Early Action Prediction	A	N/A	Early Action Prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Image Generation from Scene Graphs	A	N/A	Image Generation from Scene Graphs contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Scene Graph Classification	A	N/A	Scene Graph Classification contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Zero-Shot Semantic Segmentation	D	N/A	Zero-Shot Semantic Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Defocus Blur Detection	B	N/A	Defocus Blur Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Action Unit Detection	B	N/A	Action Unit Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Point-interactive Image Colorization	B	N/A	Point-interactive Image Colorization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Temporal Information Extraction	X	Temporal information extraction is the identification of chunks/tokens corresponding to temporal intervals, and the extraction and determination of the temporal relations between those. The entities e...	Temporal Information Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Partially Relevant Video Retrieval	A	N/A	Partially Relevant Video Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Dense Object Detection	D	N/A	Dense Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Audio-Visual Active Speaker Detection	A	N/A	Audio-Visual Active Speaker Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Graph Anomaly Detection	X	N/A	Graph Anomaly Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi-Object Tracking and Segmentation	A	Multiple object tracking and segmentation requires detecting, tracking, and segmenting objects belonging to a set of given classes. (Image and definition credit: Prototypical Cross-Attention Netw...	Multi-Object Tracking and Segmentation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Semantic Object Interaction Classification	A	N/A	Semantic Object Interaction Classification contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Event Segmentation	A	N/A	Event Segmentation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Unconditional Image Generation	B	N/A	Unconditional Image Generation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Unsupervised Facial Landmark Detection	D	N/A	Unsupervised Facial Landmark Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Unsupervised Video Summarization	A	N/A	Unsupervised Video Summarization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Multi-Hypotheses 3D Human Pose Estimation	D	N/A	Multi-Hypotheses 3D Human Pose Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Unsupervised Human Pose Estimation	D	N/A	Unsupervised Human Pose Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
3D Face Alignment	C	N/A	3D Face Alignment remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Face Hallucination	B	N/A	Face Hallucination maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Colorectal Gland Segmentation:	C	N/A	Colorectal Gland Segmentation: remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Lung Nodule Detection	C	N/A	Lung Nodule Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Molecular Property Prediction (1-shot))	X	N/A	Molecular Property Prediction (1-shot)) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Graph Property Prediction	X	N/A	Graph Property Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
Prediction	X	N/A	Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Supervised Text Retrieval	X	N/A	Supervised Text Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Human Instance Segmentation	B	Instance segmentation is the task of detecting and delineating each distinct object of interest appearing in an image. Image Credit: Deep Occlusion-Aware Instance Segmentation with Overlapping BiL...	Human Instance Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
3D Room Layouts From A Single RGB Panorama	C	N/A	3D Room Layouts From A Single RGB Panorama remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Unsupervised Saliency Detection	D	N/A	Unsupervised Saliency Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Image to sketch recognition	B	N/A	Image to sketch recognition maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Zero Shot Skeletal Action Recognition	A	N/A	Zero Shot Skeletal Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Generalized Zero Shot skeletal action recognition	A	N/A	Generalized Zero Shot skeletal action recognition contains visual evidence but needs structured language, temporal composition, retrieval, or...
Unsupervised Skeleton Based Action Recognition	A	N/A	Unsupervised Skeleton Based Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Bird's-Eye View Semantic Segmentation	C	N/A	Bird's-Eye View Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Skin Cancer Segmentation	C	N/A	Skin Cancer Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
4D Panoptic Segmentation	A	4D Panoptic Segmentation is a computer vision task that extends video panoptic segmentation to point cloud sequences. That is, given ...	4D Panoptic Segmentation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Twitter Bot Detection	X	Academic studies estimate that up to 15% of Twitter users are automated bot accounts 1. The prevalence of Twitter bots coupled with the ability of some bots to give seemingly human responses has ena...	Twitter Bot Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi-Frame Super-Resolution	A	N/A	Multi-Frame Super-Resolution contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Video Emotion Detection	A	N/A	Video Emotion Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Continuous Affect Estimation	B	N/A	Continuous Affect Estimation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Event data classification	A	N/A	Event-data classification operates on an asynchronous temporal event stream, so Difficulty A takes precedence over Difficulty C.
Open Vocabulary Action Detection	A	N/A	Open Vocabulary Action Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Target Sound Extraction	X	Target Sound Extraction is the task of extracting a sound corresponding to a given class from an audio mixture. The audio mixture may contain background noise with a relatively low amplitude compared ...	Target Sound Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Traffic Accident Detection	A	N/A	Traffic Accident Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Recognizing And Localizing Human Actions	A	N/A	Recognizing And Localizing Human Actions contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Document-level Closed Information Extraction	X	N/A	Document-level Closed Information Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi-tissue Nucleus Segmentation	C	N/A	Multi-tissue Nucleus Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Clique Prediction	X	N/A	Clique Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Polyp Segmentation	C	N/A	Polyp Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Drone Pose Estimation	D	N/A	Drone Pose Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Video Saliency Detection	A	N/A	Video Saliency Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Video Saliency Prediction	A	N/A	Video Saliency Prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Multimodal Abstractive Text Summarization	A	N/A	Multimodal Abstractive Text Summarization contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Text based Person Retrieval	A	N/A	Text based Person Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Text-based Person Retrieval with Noisy Correspondence	A	This is a benchmark about text-based person retrieval with noisy correspondence. All recorded values are the results under 20% noise rate.	Text-based Person Retrieval with Noisy Correspondence contains visual evidence but needs structured language, temporal composition, retrieval, or...

Archived task	Level	Archived description	Manual interface judgment
Semantic Retrieval	X	N/A	Semantic Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
3D Point Cloud Data Augmentation	C	N/A	3D Point Cloud Data Augmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Recognizing Emotion Cause in Conversations	X	Given an utterance U, labeled with emotion E, the task is to extract the causal spans S from the conversational history H (including utterance U) that sufficiently represent the causes of emotion E.	Recognizing Emotion Cause in Conversations has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Text-Based Stock Prediction	X	N/A	Text-Based Stock Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Audio Super-Resolution	X	Audio super-resolution, especially speech, refers to the process of reconstructing high-resolution music signals from their low-resolution counterparts. Essentially, it enhances the quality of a speec...	Audio Super-Resolution has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi Future Trajectory Prediction	A	N/A	Multi Future Trajectory Prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Grayscale Image Denoising	D	N/A	Grayscale Image Denoising is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Multi-class Anomaly Detection	D	N/A	Multi-class Anomaly Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Loop Closure Detection	A	N/A	Loop Closure Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Drug–drug Interaction Extraction	X	N/A	Drug–drug Interaction Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Online Action Detection	A	N/A	Online Action Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Chord Recognition	X	N/A	Chord Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Direction of Arrival Estimation	X	Estimating the direction-of-arrival (DOA) of a sound source from multi-channel recordings.	Direction of Arrival Estimation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Volumetric Medical Image Segmentation	C	N/A	Volumetric Medical Image Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
speech-recognition	X	N/A	speech-recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Predicate Detection	X	N/A	Predicate Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Semantic Role Labeling (predicted predicates)	X	N/A	Semantic Role Labeling (predicted predicates) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Sentence segmentation	X	N/A	Sentence segmentation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Binary Relation Extraction	X	N/A	Binary Relation Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Repetitive Action Counting	A	Repetitive action counting aims to count the number of repetitive actions in a video.	Repetitive Action Counting contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Gaze Target Estimation	B	Gaze Target Estimation refers to predicting the image 2D gaze location of a person in the image.	Gaze Target Estimation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
3D Point Cloud Matching	A	N/A	3D Point Cloud Matching contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Reflection Removal	B	Remove the spots from mirror and clear the picture	Reflection Removal maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Definition Extraction	X	N/A	Definition Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Action Spotting	A	N/A	Action Spotting contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Inverse Rendering	B	Inverse Rendering is the task of recovering the properties of a scene, such as shape, material, and lighting, from an image or a video. The goal of inverse rendering is to determine the properties...	Inverse Rendering maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
3D Panoptic Segmentation	C	N/A	3D Panoptic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Hyperspectral Semantic Segmentation	C	N/A	Hyperspectral Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Chinese Word Segmentation	X	N/A	Chinese Word Segmentation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Spindle Detection	X	N/A	Spindle Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Face Clustering	A	N/A	Face Clustering contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Event-based Optical Flow	A	N/A	Event-based optical flow is intrinsically temporal and uses asynchronous events; Difficulty A takes precedence over Difficulty C.
lidar absolute pose regression	C	N/A	lidar absolute pose regression remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.

Archived task	Level	Archived description	Manual interface judgment
Fovea Detection	C	N/A	Fovea Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Optic Disc Detection	C	Region proposal for optic disc	Optic Disc Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Object Discovery	B	Object Discovery is the task of identifying previously unseen objects. Source: Unsupervised Object Discovery and Segmentation of RGBD-images	Object Discovery maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Visual Entailment	A	N/A	Visual Entailment contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Spatio-Temporal Video Grounding	A	Spatio-temporal video grounding is a computer vision and natural language processing (NLP) task that involves linking textual descriptions to specific spatio-temporal regions or moments in a video. In...	Spatio-Temporal Video Grounding contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
3D Object Detection From Monocular Images	C	N/A	3D Object Detection From Monocular Images remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Aspect Term Extraction and Sentiment Classification	X	N/A	Aspect Term Extraction and Sentiment Classification has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Road Damage Detection	D	Road damage detection is the task of detecting damage in roads. (Image credit: Road Damage Detection And Classification In Smartphone Captured Images Using ...	Road Damage Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Texture Synthesis	B	The fundamental goal of example-based Texture Synthesis is to generate a texture, usually larger than the input, that faithfully captures all the visual characteristics of the exemplar, yet is nei...	Texture Synthesis maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Breast Cancer Histology Image Classification	C	N/A	Breast Cancer Histology Image Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
Medical Image Retrieval	A	N/A	Medical Image Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Pedestrian Trajectory Prediction	A	N/A	Pedestrian Trajectory Prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Face Morphing Attack Detection	D	N/A	Face Morphing Attack Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Voice Query Recognition	X	N/A	Voice Query Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Surgical Gesture Recognition	A	N/A	Surgical Gesture Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Affordance Detection	B	Affordance detection refers to identifying the potential action possibilities of objects in an image, which is an important ability for robot perception and manipulation. Image source: Object-Base...	Affordance Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Human-Object-interaction motion tracking	A	N/A	Human-Object-interaction motion tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Cardiac Segmentation	C	N/A	Cardiac Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Weakly supervised segmentation	D	N/A	Weakly supervised segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Sign Language Production	A	Sign Language Production (SLP) is the automatically translation from spoken language sentences into sign language sequences. Whilst Sign language Translation translates from sign to text, SLP is the o...	Sign Language Production contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Texture Image Retrieval	A	N/A	Texture Image Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Supervised Defect Detection	D	N/A	Supervised Defect Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Argument Retrieval	X	N/A	Argument Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Personality Trait Recognition	A	N/A	Personality Trait Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Semantic Image Similarity	B	N/A	Semantic Image Similarity maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Histopathological Image Classification	C	N/A	Histopathological Image Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Radar Object Detection	C	The radar object detection (ROD) task aims to classify and localize the objects in 3D purely from radar's radio frequency (RF) images.	Radar Object Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Activity Prediction	A	Predict human activities in videos	Activity Prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Satire Detection	A	N/A	Satire Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
Segmentation Of Remote Sensing Imagery	C	N/A	Segmentation Of Remote Sensing Imagery remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
3D Point Cloud Reconstruction	C	Encoding and reconstruction of 3D point clouds.	3D Point Cloud Reconstruction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Review Generation	X	N/A	Review Generation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Photo Retouching	B	N/A	Photo Retouching maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Stock Price Prediction	X	N/A	Stock Price Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Semantic Frame Parsing	X	N/A	Semantic Frame Parsing has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multilingual Named Entity Recognition	X	N/A	Multilingual Named Entity Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video Matting	A	No archived description.	Video Matting contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Camera Auto-Calibration	C	N/A	Camera Auto-Calibration remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Thermal Infrared Pedestrian Detection	C	N/A	Thermal Infrared Pedestrian Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Reading Order Detection	A	Reading order detection aims to capture the word sequence which can be naturally comprehended by human readers from visually-rich documents. It is a fundamental task for visually-rich document underst...	Reading Order Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Scientific Concept Extraction	X	N/A	Scientific Concept Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
3D Geometry Perception	C	N/A	3D Geometry Perception remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
2D Semantic Segmentation task 3 (25 classes)	B	N/A	This is ordinary 2D semantic segmentation with a dataset-specific 25-class label set; no new modality is required.
Aspect-Category-Opinion-Sentiment Quadruple Extraction	X	Aspect-Category-Opinion-Sentiment (ACOS) Quadruple Extraction is the task with the goal to extract all aspect-category-opinion-sentiment quadruples in a review sentence. (and provide full support for...	Aspect-Category-Opinion-Sentiment Quadruple Extraction has no substantive visual input, output, or visual auxiliary role in the archived task...
Multiobjective Optimization	X	Multi-objective optimization (also known as multi-objective programming, vector optimization, multicriteria optimization, multiattribute optimization or Pareto optimization) is an area of multiple cri...	Multiobjective Optimization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
3D Shape Retrieval	A	N/A	3D Shape Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Parameter Prediction	X	N/A	Parameter Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Unfairness Detection	X	N/A	Unfairness Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Extractive Summarization	X	N/A	Extractive Summarization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Key Point Matching	X	N/A	Key Point Matching has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Portrait Segmentation	B	N/A	Portrait Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Underwater Image Restoration	D	N/A	Underwater Image Restoration is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Transparent Object Depth Estimation	D	N/A	Transparent Object Depth Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Atrial Fibrillation Detection	X	N/A	Atrial Fibrillation Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Joint Event and Temporal Relation Extraction	X	N/A	Joint Event and Temporal Relation Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Aspect Category Polarity	X	N/A	Aspect Category Polarity has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Speech Denoising	X	Obtain the clean speech of the target speaker by suppressing the background noise. Recent representative github platform ClearerVoice-Studio	Speech Denoising has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Inductive Link Prediction	X	N/A	Inductive Link Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video Domain Adapation	A	N/A	Video domain adaptation remains a temporal video task; the domain-adaptation protocol does not override D.
Line Detection	B	N/A	Line Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Speech Extraction	A	N/A	Speech Extraction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
Animal Action Recognition	A	Cross-species (intra-class, inter-class) action recognition	Animal Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Multi-Animal Tracking with identification	A	N/A	Multi-Animal Tracking with identification contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Zero-shot Text Retrieval	A	N/A	Zero-shot Text Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Point Cloud Segmentation	C	3D point cloud segmentation is the process of classifying point clouds into multiple homogeneous regions, the points in the same region will have the same properties. The segmentation is challengin...	Point Cloud Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Audio Emotion Recognition	X	N/A	Audio Emotion Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
HDR Reconstruction	B	N/A	HDR Reconstruction maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Dialect Identification	X	N/A	Dialect Identification has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Blind Image Quality Assessment	B	N/A	Blind Image Quality Assessment maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Photoplethysmography (PPG) beat detection	A	N/A	Photoplethysmography (PPG) beat detection contains visual evidence but needs structured language, temporal composition, retrieval, or control...
3D Volumetric Reconstruction	C	N/A	3D volumetric reconstruction is a visual reconstruction task with an explicit volumetric output and therefore cannot be X.
Multimodal Intent Recognition	A	N/A	Multimodal Intent Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Extractive Question-Answering	X	N/A	Extractive Question-Answering has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Crop Yield Prediction	C	N/A	Crop Yield Prediction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Low-Dose X-Ray Ct Reconstruction	C	N/A	Low-Dose X-Ray Ct Reconstruction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Organ Segmentation	C	N/A	Organ Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Robust 3D Semantic Segmentation	C	N/A	Robust 3D Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Text Detection	B	Detecting the text in the image and localise it using a bounding box. The text can be in any shape and size. We need to localise all such instances of text in the entire image along with bounding box ...	Text Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Zero Shot Segmentation	D	N/A	Zero Shot Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Open Intent Detection	X	N/A	Open Intent Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
LLM-generated Text Detection	X	N/A	LLM-generated Text Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video-Adverb Retrieval (Unseen Compositions)	A	N/A	Video-Adverb Retrieval (Unseen Compositions) contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Robot Pose Estimation	C	N/A	Robot Pose Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Image to 3D	C	N/A	Image to 3D remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Audio Deepfake Detection	X	N/A	Audio Deepfake Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Camera Pose Estimation	C	Camera pose estimation is a crucial task in computer vision and robotics that involves determining the position and orientation (pose) of a camera relative to a given reference frame. This task is ess...	Camera Pose Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Attribute Mining	X	The attribute mining task assumes an open-world setting in which novel attributes and their values are extracted from free text. The task is mainly studied in the e-commerce context, where new attrib...	Attribute Mining has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Noisy Semantic Image Synthesis	B	N/A	Noisy Semantic Image Synthesis maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Key-value Pair Extraction	A	N/A	Key-value Pair Extraction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Open-Domain Subject-to-Video	A	N/A	Open-Domain Subject-to-Video contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
1 Image, 2*2 Stitchi	B	N/A	1 Image, 2*2 Stitchi maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Image to text	A	N/A	Image to text contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Medical Image Analysis	C	N/A	Medical Image Analysis remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
RGB-D Instance Segmentation	C	N/A	RGB-D Instance Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.

Archived task	Level	Archived description	Manual interface judgment
Antibody-antigen binding prediction	X	N/A	Antibody-antigen binding prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Sentence Retrieval	X	N/A	Sentence Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Nested Term Extraction	X	N/A	Nested Term Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Nested Term Recognition from Flat Supervision	X	N/A	Nested Term Recognition from Flat Supervision has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Unsupervised Image-To-Image Translation	B	Unsupervised image-to-image translation is the task of doing image-to-image translation without ground truth image-to-image pairings. (Image credit: Unpaire...	Unsupervised Image-To-Image Translation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Superpixel Image Classification	D	N/A	Superpixel Image Classification is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Semantic Textual Similarity within Bi-Encoder	X	N/A	Semantic Textual Similarity within Bi-Encoder has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Constituency Grammar Induction	X	N/A	Constituency Grammar Induction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Unsupervised Object Localization	D	N/A	Unsupervised Object Localization is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
One-Shot Object Detection	D	N/A	One-Shot Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
One-Shot Instance Segmentation	D	N/A	One-Shot Instance Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Few Shot Open Set Object Detection	D	N/A	Few Shot Open Set Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Lip to Speech Synthesis	A	Given a silent video of a speaker, generate the corresponding speech that matches the lip movements.	Lip to Speech Synthesis contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Landmark-based Lipreading	A	N/A	Landmark-based Lipreading contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Birds Eye View Object Detection	C	KITTI birds eye view detection task	Birds Eye View Object Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Monocular Cross-View Road Scene Parsing(Road)	C	N/A	Monocular Cross-View Road Scene Parsing(Road) remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
Monocular Cross-View Road Scene Parsing(Vehicle)	C	N/A	Monocular Cross-View Road Scene Parsing(Vehicle) remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as...
Image to Point Cloud Registration	A	N/A	Image to Point Cloud Registration contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Self-supervised Video Retrieval	A	N/A	Self-supervised Video Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Length-of-Stay prediction	X	N/A	Length-of-Stay prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi-label Image Recognition with Partial Labels	D	N/A	Multi-label Image Recognition with Partial Labels is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already...
Unsupervised semantic parsing	A	N/A	Unsupervised semantic parsing contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Bird Species Classification With Audio-Visual Data	C	N/A	Bird Species Classification With Audio-Visual Data remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added...
Zero-shot Relation Triplet Extraction	X	N/A	Zero-shot Relation Triplet Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video Story QA	A	N/A	Video Story QA contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Handwritten Chinese Text Recognition	B	N/A	Handwritten Chinese Text Recognition maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Distractor Generation	A	N/A	Distractor Generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Few Shot Temporal Action Localization	A	N/A	Few Shot Temporal Action Localization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Semi-Supervised Action Detection	A	N/A	Semi-Supervised Action Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Zero-Shot Action Detection	A	N/A	Zero-Shot Action Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
One-stage Anchor-free Oriented Object Detection	D	N/A	One-stage Anchor-free Oriented Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already...
Subjectivity Analysis	X	A related task to sentiment analysis is the subjectivity analysis with the goal of labeling an opinion as either subjective or objective.	Subjectivity Analysis has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Traffic Object Detection	D	N/A	Traffic Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Zero-Shot Human-Object Interaction Detection	A	N/A	The zero-shot modifier is only a protocol; human-object interaction detection has structured relational output requiring D.
Generic Event Boundary Detection	A	N/A	Generic Event Boundary Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...

Archived task	Level	Archived description	Manual interface judgment
Speaker-Specific Lip to Speech Synthesis	A	N/A	Speaker-Specific Lip to Speech Synthesis contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Group Activity Recognition	A	N/A	Group Activity Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Fine-Grained Action Detection	A	N/A	Fine-Grained Action Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Image Rescaling	B	N/A	Image Rescaling maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
3D Canonical Hand Pose Estimation	C	Image: Lin et al	3D Canonical Hand Pose Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
3D Shape Reconstruction From A Single 2D Image	C	Image: Liao et al	3D Shape Reconstruction From A Single 2D Image remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
3D Car Instance Understanding	C	N/A	3D Car Instance Understanding remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Highlight Detection	A	No archived description.	Highlight Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Continual Semantic Segmentation	D	Continual learning in semantic segmentation.	Continual Semantic Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Weakly-supervised 3D Human Pose Estimation	D	N/A	Weakly-supervised 3D Human Pose Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Unsupervised 3D Human Pose Estimation	D	N/A	Unsupervised 3D Human Pose Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Disguised Face Verification	D	N/A	Disguised Face Verification is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Age-Invariant Face Recognition	D	N/A	Age-Invariant Face Recognition is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Dichotomous Image Segmentation	B	N/A	Dichotomous Image Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Lung Nodule Segmentation	C	N/A	Lung Nodule Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Graph structure learning	X	N/A	Graph structure learning has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Ancestor-descendant prediction	X	Given two entities, make a binary prediction if they have ancestor-descendant relationship, based on existing and missing hierarchical edges in the graph.	Ancestor-descendant prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Aspect-Based Sentiment Analysis	X	N/A	Aspect-Based Sentiment Analysis has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Aspect-oriented Opinion Extraction	X	N/A	Aspect-oriented Opinion Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Curved Text Detection	B	N/A	Curved Text Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Brain Image Segmentation	C	N/A	Brain Image Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Link Sign Prediction	X	N/A	Link Sign Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Resynthesis	X	N/A	Resynthesis has no substantive visual input, output, or visual auxiliary role in the archived task definition.
One-Shot Segmentation	D	(Image credit: One-Shot Learning for Semantic Segmentation)	One-Shot Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Self-supervised Skeleton-based Action Recognition	A	N/A	Self-supervised Skeleton-based Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or...
Zero-Shot Video-Audio Retrieval	A	N/A	Zero-Shot Video-Audio Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
3D Multi-Person Pose Estimation (absolute)	C	This task aims to solve absolute 3D multi-person pose Estimation (camera-centric coordinates). No ground truth human bounding box and human root joint coordinates are used during testing stage. ¡sp...	3D Multi-Person Pose Estimation (absolute) remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Prosody Prediction	X	N/A	Prosody Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Pneumonia Detection	C	N/A	Pneumonia Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Table-based Fact Verification	X	Verifying facts given semi-structured data.	Table-based Fact Verification has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Cross-domain 3D Human Pose Estimation	D	N/A	Cross-domain 3D Human Pose Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
HD semantic map learning	A	The goal of task is to generate map elements in a vectorized form using data from onboard sensors, e.g., RGB cameras and/or LiDARs. These map elements include but are not limited to : Road boundaries,...	HD semantic map learning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Motion Detection	A	Motion Detection is a process to detect the presence of any moving entity in an area of interest. Motion Detection is of great importance due to its application in various areas such as surveillan...	Motion Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
satellite image super-resolution	C	N/A	satellite image super-resolution remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Arrhythmia Detection	X	N/A	Arrhythmia Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Weakly Supervised Action Segmentation (Transcript)	A	N/A	Weakly Supervised Action Segmentation (Transcript) contains visual evidence but needs structured language, temporal composition, retrieval, or...
Multiple Object Tracking with Transformer	A	N/A	Multiple Object Tracking with Transformer contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Image Retrieval with Multi-Modal Query	A	The problem of retrieving images from a database based on a multi-modal (image-text) query. Specifically, the query text prompts some modification in the query image and the task is to retrieve image...	Image Retrieval with Multi-Modal Query contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Chinese Semantic Role Labeling	X	N/A	Chinese Semantic Role Labeling has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Zero-shot dense video captioning	A	N/A	Zero-shot dense video captioning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Text-Line Extraction	B	N/A	Text-Line Extraction maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
3D Object Reconstruction From A Single Image	C	N/A	3D Object Reconstruction From A Single Image remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
Scanpath prediction	A	Learning to Predict Sequences of Human Fixations.	Scanpath prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Document Enhancement	B	N/A	Document Enhancement maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Camouflaged Object Segmentation with a Single Task-generic Prompt	A	N/A	Camouflaged Object Segmentation with a Single Task-generic Prompt contains visual evidence but needs structured language, temporal composition,...
Fine-Grained Urban Flow Inference	A	Fine-grained urban flow inference (FUFI) aims to infer the fine-grained urban flow map from the coarse-grained one.	Fine-Grained Urban Flow Inference contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Co-Salient Object Detection	B	N/A	Co-Salient Object Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Single-object colocalization	B	N/A	Single-object colocalization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Action Assessment	A	N/A	Action Assessment contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Dialog Relation Extraction	X	N/A	Dialog Relation Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Zero-Shot Transfer 3D Point Cloud Classification	C	N/A	Zero-Shot Transfer 3D Point Cloud Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as...
3D Point Cloud Linear Classification	C	Training a linear classifier(e.g. SVM) on the embeddings/representations of 3D point clouds. The embeddings/representations are usually trained in an unsupervised manner.	3D Point Cloud Linear Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Training-free 3D Point Cloud Classification	C	Evaluation on target datasets for 3D Point Cloud Classification without any training	Training-free 3D Point Cloud Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
3D Parameter-Efficient Fine-Tuning for Classification	C	N/A	3D Parameter-Efficient Fine-Tuning for Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be...
Conditional Text-to-Image Synthesis	A	N/A	Conditional Text-to-Image Synthesis contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Query-Based Extractive Summarization	X	N/A	Query-Based Extractive Summarization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Inductive logic programming	X	N/A	Inductive logic programming has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Single-Image Portrait Relighting	B	N/A	Single-Image Portrait Relighting maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Fact-based Text Editing	X	N/A	Fact-based Text Editing has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi-domain Dialogue State Tracking	X	N/A	Multi-domain Dialogue State Tracking has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Iris Segmentation	C	N/A	Iris Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Foggy Scene Segmentation	D	N/A	Foggy Scene Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Source Free Object Detection	D	N/A	Source Free Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Unsupervised Few-Shot Image Classification	D	N/A	Unsupervised Few-Shot Image Classification is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
zero-shot anomaly detection	D	N/A	zero-shot anomaly detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Scene Classification (unified classes)	D	N/A	Scene Classification (unified classes) is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Robust Semi-Supervised RGBD Semantic Segmentation	C	N/A	Robust Semi-Supervised RGBD Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added...

Archived task	Level	Archived description	Manual interface judgment
Generative 3D Object Classification	C	The task of generative 3D object classification involves prompting the model to generate the object type from its point cloud, distinguishing it from discriminative models that directly classify objec...	Generative 3D Object Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Single Image Haze Removal	B	N/A	Single Image Haze Removal maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Depth Map Super-Resolution	B	N/A	Depth Map Super-Resolution maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Prediction Intervals	X	A prediction interval is an estimate of an interval in which a future observation will fall, with a certain probability, given what has already been observed. Prediction intervals are often used in re...	Prediction Intervals has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Zero-shot Text to Audio Retrieval	X	N/A	Zero-shot Text to Audio Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Diffeomorphic Medical Image Registration	A	N/A	Diffeomorphic Medical Image Registration contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Liver Segmentation	C	N/A	Liver Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Extractive Document Summarization	X	N/A	Extractive Document Summarization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Sequential sentence segmentation	X	N/A	Sequential sentence segmentation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Sentence Compression	X	Sentence Compression is the task of reducing the length of text by removing non-essential content while preserving important facts and grammaticality.	Sentence Compression has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Unseen Object Instance Segmentation	D	N/A	Unseen Object Instance Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Nonhomogeneous Image Dehazing	B	N/A	Nonhomogeneous Image Dehazing maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Video Forensics	A	N/A	Video Forensics contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Sleep spindles detection	X	N/A	Sleep spindles detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Product Categorization	X	N/A	Product Categorization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Handwritten Word Segmentation	B	N/A	Handwritten Word Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Atomic action recognition	A	N/A	Atomic action recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Semantic Composition	X	Understanding the meaning of text by composing the meanings of the individual words in the text.	Semantic Composition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Part-aware Panoptic Segmentation	B	Panoptic segmentation with part-aware predictions.	Part-aware Panoptic Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Single Image Desnowing	B	N/A	Single Image Desnowing maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Fish Detection	D	N/A	Fish Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Heterogeneous Face Recognition	D	N/A	Heterogeneous Face Recognition is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Diabetic Retinopathy Detection	C	N/A	Diabetic Retinopathy Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Paper generation (Title-to-abstract)	X	N/A	Paper generation (Title-to-abstract) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Paper generation (abstract-to-conclusion)	X	N/A	Paper generation (abstract-to-conclusion) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multi-View 3D Reconstruction	A	N/A	Multi-View 3D Reconstruction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Face Age Editing	B	N/A	Face Age Editing maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Semantic entity labeling	A	- One of Form Understanding task (Word grouping, Semantic entity labeling, Entity linking) - Classifying entities into one of four pre-defined categories: question, answer, header and, other. cite...	Semantic entity labeling contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
PointGoal Navigation	A	N/A	PointGoal Navigation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
3D Shape Recognition	C	N/A	3D Shape Recognition remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Image Harmonization	B	N/A	Image Harmonization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Motion Captioning	A	Generating textual description for human motion.	Motion Captioning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
Active Speaker Detection	A	N/A	Active Speaker Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Deception Detection	A	N/A	Deception Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
K-complex detection	X	N/A	K-complex detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Depth Aleatoric Uncertainty Estimation	B	N/A	Depth Aleatoric Uncertainty Estimation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Head Detection	B	N/A	Head Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Selection bias	X	N/A	Selection bias has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Thermal Infrared Object Tracking	A	N/A	Thermal Infrared Object Tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Optic Cup Segmentation	C	Optic cup segmentation, concentric with optic disc, useful for glaucoma management (ophthalmology)	Optic Cup Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Classify 3D Point Clouds	C	N/A	Classify 3D Point Clouds remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Micro Expression Recognition	A	N/A	Micro Expression Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
3D dense captioning	A	N/A	3D dense captioning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Protein Structure Prediction	X	Image credit: FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours	Protein Structure Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Vector Graphics	A	N/A	Vector Graphics contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
self-supervised scene text recognition	A	scene text recognition task self-supervised scene text recognition task text segmentation task text image super-resolution task	self-supervised scene text recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Counterfactual Inference	X	N/A	Counterfactual Inference has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Factual probe	X	N/A	Factual probe has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Semi-supervised Medical Image Segmentation	C	N/A	Semi-supervised Medical Image Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Object Reconstruction	C	N/A	Object Reconstruction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Visual Social Relationship Recognition	A	N/A	Visual Social Relationship Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
EMG Gesture Recognition	X	Electromyographic Gesture Recognition	EMG Gesture Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Medical Procedure	X	Predicting medical procedures performed during a hospital admission.	Medical Procedure has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Open Vocabulary Action Recognition	A	N/A	Open Vocabulary Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Electron Microscopy Image Segmentation	C	N/A	Electron Microscopy Image Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Pulmorary Vessel Segmentation	C	N/A	Pulmorary Vessel Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Semi-supervised Anomaly Detection	D	N/A	Semi-supervised Anomaly Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Commonsense Knowledge Base Construction	X	N/A	Commonsense Knowledge Base Construction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Text-to-video search	A	N/A	Text-to-video search contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Self-Supervised Anomaly Detection	D	N/A	Self-Supervised Anomaly Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Hope Speech Detection	X	Detecting speech associated with positive, uplifting, promise, potential, support, reassurance, suggestions, or inspiration.	Hope Speech Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Speech Synthesis - Gujarati	X	N/A	Speech Synthesis - Gujarati has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Citation Prediction	X	N/A	Citation Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Tweet Retrieval	X	N/A	Tweet Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
News Retrieval	X	N/A	News Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
severity prediction	C	N/A	severity prediction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Anomaly Segmentation	B	N/A	Anomaly Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Zero-Shot Audio Retrieval	X	N/A	Zero-Shot Audio Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
drone-based object tracking	A	drone-based object tracking	drone-based object tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Partial Point Cloud Matching	A	N/A	Partial Point Cloud Matching contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Span-Extraction MRC	X	N/A	Span-Extraction MRC has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Graph Reconstruction	X	N/A	Graph Reconstruction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Dialog Act Classification	X	N/A	Dialog Act Classification has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Time-to-Event Prediction	X	N/A	Time-to-Event Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Joint Vertebrae Identification And Localization In Spinal Ct Images	C	N/A	Joint Vertebrae Identification And Localization In Spinal Ct Images remains a visual mapping, but requires an aligned sensor/3D/medical...
Model extraction	X	Model extraction attacks, aka model stealing attacks, are used to extract the parameters from the target model. Ideally, the adversary will be able to steal and replicate a model that will have a very...	Model extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Single-Image-Based Hdr Reconstruction	B	N/A	Single-Image-Based Hdr Reconstruction maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Sentiment Analysis (Product + User)	X	N/A	Sentiment Analysis (Product + User) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Indoor Scene Synthesis	B	N/A	Indoor Scene Synthesis maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Time Series Alignment	X	N/A	Time Series Alignment has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Survival Prediction	X	N/A	Survival Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Emotional Speech Synthesis	X	N/A	Emotional Speech Synthesis has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Image Similarity Detection	B	A fundamental computer vision task to determine whether a part of an image has been copied from another image. Description from: The 2021 Image Similarity Dataset and Challenge	Image Similarity Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Image Similarity Search	A	Image credit: The 2021 Image Similarity Dataset and Challenge	Image Similarity Search contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
3D Multi-Person Mesh Recovery	C	N/A	3D Multi-Person Mesh Recovery remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Multimodal Text Prediction	A	Multimodal text prediction is a type of natural language processing that involves predicting the next word or sequence of words in a sentence, given multiple modalities or types of input. In tradi...	Multimodal Text Prediction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Active Speaker Localization	A	Active Speaker Localization (ASL) is the process of spatially localizing an active speaker (talker) in an environment using either audio, vision or both.	Active Speaker Localization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Factual Visual Question Answering	A	N/A	Factual Visual Question Answering contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Image Matching	A	N/A	Image Matching contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Vehicle Speed Estimation	A	Vehicle speed estimation is the task of detecting and tracking vehicles whose real-world speeds are then estimated. The task is usually evaluated with recall and precision of the detected vehicle trac...	Vehicle Speed Estimation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Fine-Grained Vehicle Classification	D	N/A	Fine-Grained Vehicle Classification is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
QRS Complex Detection	X	N/A	QRS Complex Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video Visual Relation Detection	A	N/A	Video Visual Relation Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Video Visual Relation Tagging	A	N/A	Video Visual Relation Tagging contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Video scene graph generation	A	N/A	Video scene graph generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Activity Recognition In Videos	A	N/A	Activity Recognition In Videos contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Low-light Pedestrian Detection	D	N/A	Low-light Pedestrian Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Point Cloud Quality Assessment	C	### Background A large and dense collection of points in three-dimensional space, collected by sensors such as LiDAR, is known as a point cloud. Points in the point cloud consist of geometric propert...	Point Cloud Quality Assessment remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.

Archived task	Level	Archived description	Manual interface judgment
Aspect Sentiment Triplet Extraction	X	N/A	Aspect Sentiment Triplet Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Document-level Event Extraction	X	N/A	Document-level Event Extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
New Product Sales Forecasting	A	N/A	New Product Sales Forecasting contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Classification of toxic, engaging, fact-claiming comments	X	N/A	Classification of toxic, engaging, fact-claiming comments has no substantive visual input, output, or visual auxiliary role in the archived task...
Zero-Shot Cross-Lingual Visual Reasoning	A	N/A	Zero-Shot Cross-Lingual Visual Reasoning contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Max-Shot Cross-Lingual Visual Reasoning	A	N/A	Max-Shot Cross-Lingual Visual Reasoning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Moving Object Detection	A	N/A	Moving Object Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
2D Semantic Segmentation task 1 (8 classes)	B	N/A	2D Semantic Segmentation task 1 (8 classes) maps visual evidence to a dense or categorical visual output and fits the existing interface with a...
3D Geometry Prediction	C	N/A	3D Geometry Prediction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Audio Effects Modeling	X	Modeling of audio effects such as reverberation, compression, distortion, etc.	Audio Effects Modeling has no substantive visual input, output, or visual auxiliary role in the archived task definition.
3D Dense Shape Correspondence	A	N/A	3D Dense Shape Correspondence contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Acoustic echo cancellation	X	N/A	Acoustic echo cancellation has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Emotional Dialogue Acts	X	N/A	Emotional Dialogue Acts has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video Synchronization	A	N/A	Video Synchronization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Medical Image Denoising	C	Image credit: Learning Medical Image Denoising with Deep Dynamic Residual Attention Network	Medical Image Denoising remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Transparent Object Detection	B	Detecting transparent objects in 2D or 3D	Transparent Object Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Seismic Detection	X	N/A	Seismic Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Dark Humor Detection	X	N/A	Dark Humor Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Similarities Abstraction	A	N/A	Similarities Abstraction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Muscle Tendon Junction Identification	C	N/A	Muscle Tendon Junction Identification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
3D human pose and shape estimation	C	Estimate 3D human pose and shape (e.g. SMPL) from images	3D human pose and shape estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Infrared image super-resolution	C	Aims at upsampling the IR image and create the high resolution image with help of a low resolution image.	Infrared image super-resolution remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Rgb-T Tracking	A	N/A	Rgb-T Tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Type prediction	X	N/A	Type prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Kinematic Based Workflow Recognition	A	N/A	Kinematic Based Workflow Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Tomographic Reconstructions	C	N/A	Tomographic Reconstructions remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
COVID-19 Tracking	C	N/A	COVID-19 Tracking remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Temporal Metadata Manipulation Detection	A	N/A	Temporal Metadata Manipulation Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control...
3D Anomaly Detection and Segmentation	C	N/A	3D Anomaly Detection and Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Pancreas Segmentation	C	N/A	Pancreas Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Event-based Motion Estimation	C	N/A	Event-based Motion Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Claim Extraction with Stance Classification (CESC)	X	N/A	Claim Extraction with Stance Classification (CESC) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multiview Contextual Commonsense Inference	X	N/A	Multiview Contextual Commonsense Inference has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Hyperspectral Image Super-Resolution	C	N/A	Hyperspectral Image Super-Resolution remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.

Archived task	Level	Archived description	Manual interface judgment
Mistake Detection	A	Mistakes are natural occurrences in many tasks and an opportunity for an AR assistant to provide help. Identifying such mistakes requires modelling procedural knowledge and retaining long-range sequen...	Mistake Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
AbbreviationDetection	X	N/A	AbbreviationDetection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Object Categorization	B	Object categorization identifies which label, from a given set, best corresponds to an image region defined by an input image and bounding box.	Object Categorization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Retinal OCT Disease Classification	C	N/A	Retinal OCT Disease Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Data Visualization	X	N/A	Data Visualization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Copy Detection	A	N/A	Copy Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is the...
Audio-visual Question Answering	A	N/A	Audio-visual Question Answering contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Weakly-supervised Video Anomaly Detection	A	N/A	Weakly-supervised Video Anomaly Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Burst Image Super-Resolution	A	N/A	Burst Image Super-Resolution contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Multi-Oriented Scene Text Detection	B	N/A	Multi-Oriented Scene Text Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Transparent objects	D	N/A	Transparent objects is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Dialogue Safety Prediction	X	N/A	Dialogue Safety Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Amodal Instance Segmentation	B	N/A	Amodal Instance Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Variable Detection	X	N/A	Variable Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Spam detection	X	N/A	Spam detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Brain Tumor Classification	C	N/A	Brain Tumor Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Motion Generation	A	N/A	Motion Generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Electrocardiography (ECG)	X	N/A	Electrocardiography (ECG) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Burst Image Reconstruction	A	N/A	Burst Image Reconstruction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Training-free Object Counting	D	N/A	Training-free Object Counting is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Skin Lesion Segmentation	C	N/A	Skin Lesion Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Ad-Hoc Information Retrieval	X	Ad-hoc information retrieval refers to the task of returning information resources related to a user query formulated in natural language.	Ad-Hoc Information Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Semantic Communication	A	N/A	Semantic Communication contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Drug Response Prediction	X	N/A	Drug Response Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Twitter Event Detection	X	Detection of worldwide events from categories like Sports, Politics, Entertainment, Science & Technology, etc. by analyzing Twitter Tweets.	Twitter Event Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Punctuation Restoration	X	Punctuation Restoration	Punctuation Restoration has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Zero-shot 3D classification	C	N/A	Zero-shot 3D classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
3D Object Captioning	A	3D object captioning involves generating a natural language description of an object, given its point cloud representation.	3D Object Captioning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Emotional Intelligence	X	Emotional Intelligence (EI) is a measure of "The ability to monitor one's own and others' feelings, to discriminate among them, and to use this information to guide one's thinking and action." (Salove...	Emotional Intelligence has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Depth-aware Video Panoptic Segmentation	A	N/A	Depth-aware Video Panoptic Segmentation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Connectivity Estimation	X	N/A	Connectivity Estimation has no substantive visual input, output, or visual auxiliary role in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
Audio Synthesis	A	N/A	Audio Synthesis contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Semi-supervised Medical Image Classification	C	N/A	Semi-supervised Medical Image Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
Fake Image Attribution	B	N/A	Fake Image Attribution maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Group Anomaly Detection	A	N/A	Group Anomaly Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
legal outcome extraction	X	N/A	legal outcome extraction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Hyperspectral Image Denoising	C	N/A	Hyperspectral Image Denoising remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Video-to-image Affordance Grounding	A	Given a demonstration video V and a target image I, the goal of video-to-image affordance grounding predict an affordance heatmap over the target image according to the hand-interacted region in the v...	Video-to-image Affordance Grounding contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Robust 3D Object Detection	C	N/A	Robust 3D Object Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
audio-visual event localization	A	N/A	audio-visual event localization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Table Retrieval	X	N/A	Table Retrieval has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Preference Mapping	A	N/A	Preference Mapping contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
3D Point Cloud Interpolation	A	Point cloud interpolation is a fundamental problem for 3D computer vision. Given a low temporal resolution (frame rate) point cloud sequence, the target of interpolation is to generate a smooth point ...	3D Point Cloud Interpolation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Video to Text Retrieval	A	N/A	Video to Text Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Irregular Text Recognition	B	N/A	Irregular Text Recognition maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Protein Function Prediction	X	For GO terms prediction, given the specific function prediction instruction and a protein sequence, models characterize the protein functions using the GO terms presented in three different domains (c...	Protein Function Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Occluded Face Detection	D	N/A	Occluded Face Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Video-based Generative Performance Benchmarking	A	The benchmark evaluates a generative Video Conversational Model and covers five key aspects: - Correctness of Information - Detailed Orientation - Contextual Understanding - Temporal Understandi...	Video-based Generative Performance Benchmarking contains visual evidence but needs structured language, temporal composition, retrieval, or...
Image Shadow Removal	B	Merge with the Shadow Removal	Image Shadow Removal maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Image-guided Story Ending Generation	A	N/A	Image-guided Story Ending Generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Zero-shot Composed Person Retrieval	A	N/A	Zero-shot Composed Person Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
visual instruction following	A	N/A	visual instruction following contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Explanatory Visual Question Answering	A	Explanatory Visual Question Answering (EVQA) requires answering visual questions and generating multimodal explanations for the reasoning processes.	Explanatory Visual Question Answering contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Music Performance Rendering	X	N/A	Music Performance Rendering has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Low-light Image Deblurring and Enhancement	A	N/A	Low-light Image Deblurring and Enhancement contains visual evidence but needs structured language, temporal composition, retrieval, or control...
X-ray PDF regression	C	N/A	X-ray PDF regression remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
named-entity-recognition	X	N/A	named-entity-recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multispectral Image Super-resolution	C	N/A	Multispectral Image Super-resolution remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
zero-shot long video question answering	A	N/A	zero-shot long video question answering contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
3D visual grounding	A	N/A	3D visual grounding contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Counterfactual Reasoning	A	N/A	Counterfactual Reasoning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Flood Inundation Mapping	C	N/A	Flood Inundation Mapping remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Articulated Object modelling	C	N/A	Articulated Object modelling remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.

Archived task	Level	Archived description	Manual interface judgment
Action Classification (1-shot)	X	N/A	Action Classification (1-shot) has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Fault localization	X	N/A	Fault localization has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Detecting Image Manipulation	B	N/A	Detecting Image Manipulation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Garment Reconstruction	C	N/A	Garment Reconstruction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Marine Animal Segmentation	D	N/A	Marine Animal Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Mirror Detection	B	N/A	Mirror Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Keypoint detection and image matching	A	keypoint detection in retinal images followed by image registration	Keypoint detection and image matching contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Causal Discovery in Video Reasoning	A	N/A	Causal Discovery in Video Reasoning contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Audio Denoising	X	N/A	Audio Denoising has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Video Deblurring	A	N/A	Video Deblurring contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Video deraining	A	N/A	Video deraining contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Music Genre Recognition	X	Recognizing the genre (e.g. rock, pop, jazz, etc.) of a piece of music.	Music Genre Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Text-to-Video Editing	A	N/A	Text-to-Video Editing contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Outdoor Localization	A	N/A	Outdoor Localization contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Synthetic Image Detection	B	Identify if the image is real or generated/manipulated by any generative models (GAN or Diffusion).	Synthetic Image Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Malaria Falciparum Detection	C	N/A	Malaria Falciparum Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Text within image generation	A	N/A	Text within image generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Water3D_Long	C	N/A	Water3D_Long remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Sand3D_Long	C	N/A	Sand3D_Long remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Plasticine3D	C	N/A	Plasticine3D remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
3D geometry	C	N/A	3D geometry remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Eyeblink detection	A	Localize eyeblinks in videos.	Eyeblink detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Audio-Video Question Answering (AVQA)	A	N/A	Audio-Video Question Answering (AVQA) contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Damaged Building Detection	C	N/A	Damaged Building Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Foreground Segmentation	B	N/A	Foreground Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Image Editing	B	N/A	Image Editing maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
1 Image, 2*2 Stitching	B	N/A	1 Image, 2*2 Stitching maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Malicious Detection	D	N/A	Malicious Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Iloko Speech Recognition	X	N/A	Iloko Speech Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
PD-L1 Tumor Proportion Score Prediction	C	N/A	PD-L1 Tumor Proportion Score Prediction remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
HairColor/Bias-conflicting	D	N/A	HairColor/Bias-conflicting is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
HeavyMakeup/Bias-conflicting	D	N/A	HeavyMakeup/Bias-conflicting is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
JPEG Decompression	B	Image credit: Palette: Image-to-Image Diffusion Models	JPEG Decompression maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Image Classification with Differential Privacy	D	N/A	Image Classification with Differential Privacy is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already...
Zero-Shot Transfer Image Classification (CN)	D	N/A	Zero-Shot Transfer Image Classification (CN) is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Unsupervised face recognition	D	N/A	Unsupervised face recognition is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Training-free 3D Part Segmentation	C	Evaluation on target datasets for 3D Part Segmentation without any training	Training-free 3D Part Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.

Archived task	Level	Archived description	Manual interface judgment
Weakly-supervised instance segmentation	D	N/A	Weakly-supervised instance segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Image-level Supervised Instance Segmentation	D	N/A	Image-level Supervised Instance Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Point-Supervised Instance Segmentation	D	N/A	Point-Supervised Instance Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Active Object Detection	B	N/A	Active Object Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Box-supervised Instance Segmentation	D	N/A	Box-supervised Instance Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Activeness Detection	B	N/A	Activeness Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Generalized Zero-Shot Object Detection	D	N/A	Generalized Zero-Shot Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
3D Single Object Tracking	A	N/A	3D Single Object Tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Open Set Action Recognition	A	N/A	Open Set Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Multi-Label Classification Of Biomedical Texts	X	N/A	Multi-Label Classification Of Biomedical Texts has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Bidirectional Relationship Classification	A	N/A	Bidirectional Relationship Classification contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Unbiased Scene Graph Generation	A	N/A	Unbiased Scene Graph Generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Smile Recognition	B	N/A	Smile Recognition maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output adjustment.
Unsupervised Keypoint Estimation	D	N/A	Unsupervised Keypoint Estimation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Image Comprehension	A	N/A	Image Comprehension contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode...
Clinical Assertion Status Detection	X	Classifying the assertions made on given medical concepts as being present, absent, or possible in the patient, conditionally present in the patient under certain circumstances, hypothetically present...	Clinical Assertion Status Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
GZSL Video Classification	A	N/A	GZSL Video Classification contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
ZSL Video Classification	A	N/A	ZSL Video Classification contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Cross-Part Crowd Counting	D	N/A	Cross-Part Crowd Counting is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Zero-Shot Single Object Tracking	A	N/A	Zero-Shot Single Object Tracking contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Drivable Area Detection	B	The drivable area detection is a subset topic of object detection. The model marks the safe and legal roads for regular driving in color blocks shaped by area.	Drivable Area Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Multiple Object Track and Segmentation	A	N/A	Multiple Object Track and Segmentation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Amodal Panoptic Segmentation	B	N/A	Amodal Panoptic Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Affordance Recognition	B	Affordance recognition from Human-Object Interaction	Affordance Recognition maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Human-Object Interaction Concept Discovery	A	Discovering the reasonable HOI concepts/categories from known categories and their instances. Actually, it is also a matrix (verb-object matrix) complementation problem.	Human-Object Interaction Concept Discovery contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Phoneme Recognition	X	N/A	Phoneme Recognition has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Jpeg Compression Artifact Reduction	B	N/A	Jpeg Compression Artifact Reduction maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Multiview Gait Recognition	A	N/A	Multiview Gait Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Occluded 3D Object Symmetry Detection	C	N/A	Occluded 3D Object Symmetry Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Vehicle Key-Point and Orientation Estimation	C	N/A	Vehicle Key-Point and Orientation Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
Car Pose Estimation	C	N/A	Car Pose Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Semi-Supervised Image Classification (Cold Start)	D	N/A	Semi-Supervised Image Classification (Cold Start) is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already...
Neural Network Compression	X	N/A	Neural Network Compression has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Image Classification with Human Noise	D	N/A	Image Classification with Human Noise is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.

Archived task	Level	Archived description	Manual interface judgment
Supervised Image Retrieval	A	N/A	Supervised Image Retrieval contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Open Vocabulary Panoptic Segmentation	D	N/A	Open Vocabulary Panoptic Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Speech Prompted Semantic Segmentation	C	N/A	Speech Prompted Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Sound Prompted Semantic Segmentation	C	N/A	Sound Prompted Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Open-Vocabulary Semantic Segmentation	D	N/A	Open-Vocabulary Semantic Segmentation is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Root Joint Localization	B	N/A	Root Joint Localization maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
3D Human Pose Estimation in Limited Data	D	N/A	3D Human Pose Estimation in Limited Data is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
3D Semantic Instance Segmentation	C	N/A	3D Semantic Instance Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Interactive 3D Instance Segmentation	C	N/A	Interactive 3D Instance Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Unsupervised 3D Semantic Segmentation	C	N/A	Unsupervised 3D Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM...
Interactive 3D Instance Segmentation -Trained on Scannet40 - Evaluated on Scannet40	C	N/A	Interactive 3D Instance Segmentation -Trained on Scannet40 - Evaluated on Scannet40 remains a visual mapping, but requires an aligned...
Semantic Contour Prediction	B	N/A	Semantic Contour Prediction maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Object Skeleton Detection	B	N/A	Object Skeleton Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Hyperspectral Image Inpainting	C	N/A	Hyperspectral Image Inpainting remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Semantic Part Detection	B	N/A	Semantic Part Detection maps visual evidence to a dense or categorical visual output and fits the existing interface with a small label/output...
Age/Bias-conflicting	D	N/A	Age/Bias-conflicting is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Race/Bias-conflicting	D	N/A	Race/Bias-conflicting is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Participant Intervention Comparison Outcome Extraction	X	PICO recognition is an information extraction task for identifying Participant, Intervention, Comparator, and Outcome (PICO elements) information from clinical literature.	Participant Intervention Comparison Outcome Extraction has no substantive visual input, output, or visual auxiliary role in the archived task...
Pose-Based Human Instance Segmentation	B	N/A	Pose-Based Human Instance Segmentation maps visual evidence to a dense or categorical visual output and fits the existing interface with a small...
Geometric Matching	A	N/A	Geometric Matching contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Class-agnostic Object Detection	D	N/A	Class-agnostic Object Detection is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Relationship Detection	A	N/A	Relationship Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Few-shot 3D Point Cloud Semantic Segmentation	C	N/A	Few-shot 3D Point Cloud Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
Photo to Rest Generalization	D	N/A	Photo to Rest Generalization is mainly a domain, supervision, robustness, or evaluation variant of an interface WSM already supports.
Gene Interaction Prediction	X	N/A	Gene Interaction Prediction has no substantive visual input, output, or visual auxiliary role in the archived task definition.
3D Multi-Person Pose Estimation (root-relative)	C	This task aims to solve root-relative 3D multi-person pose estimation (person-centric coordinate system). No ground truth human bounding box and human root joint coordinates are used during testing st...	3D Multi-Person Pose Estimation (root-relative) remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as...
Unsupervised 3D Multi-Person Pose Estimation	C	N/A	Unsupervised 3D Multi-Person Pose Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a...
General Action Video Anomaly Detection	A	N/A	General Action Video Anomaly Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
One-Shot 3D Action Recognition	A	N/A	One-Shot 3D Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Self-Supervised Human Action Recognition	A	N/A	Self-Supervised Human Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control...
Few-Shot Skeleton-Based Action Recognition	A	N/A	Few-Shot Skeleton-Based Action Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control...
One-shot visual object segmentation	A	N/A	One-shot visual object segmentation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Multi-view 3D Human Pose Estimation	C	N/A	Multi-view 3D Human Pose Estimation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Online Vectorized HD Map Construction	A	N/A	Online Vectorized HD Map Construction contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...

Archived task	Level	Archived description	Manual interface judgment
Sleep Arousal Detection	X	N/A	Sleep Arousal Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Multimodal Sleep Stage Detection	X	N/A	Multimodal Sleep Stage Detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Myocardial infarction detection	X	N/A	Myocardial infarction detection has no substantive visual input, output, or visual auxiliary role in the archived task definition.
Lung Nodule 3D Classification	C	N/A	Lung Nodule 3D Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Lung Nodule 3D Detection	C	N/A	Lung Nodule 3D Detection remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Multi-view Subspace Clustering	A	N/A	Multi-view Subspace Clustering contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Video-to-Video Synthesis	A	N/A	Video-to-Video Synthesis contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse...
Real-Time 3D Semantic Segmentation	C	N/A	Real-Time 3D Semantic Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Lidar Scene Completion	C	Obtaining dense scene representation from a sparse lidar point cloud.	Lidar Scene Completion remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Text-to-3D-Human Generation	A	3D avatars generation from text prompts	Text-to-3D-Human Generation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Visual Question Answering (VQA) Split A	A	N/A	Visual Question Answering (VQA) Split A contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Visual Question Answering (VQA) Split B	A	N/A	Visual Question Answering (VQA) Split B contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Drone navigation	A	N/A	Drone navigation contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's sparse mode is...
Weakly Supervised Action Segmentation (Action Set))	A	N/A	Weakly Supervised Action Segmentation (Action Set)) contains visual evidence but needs structured language, temporal composition, retrieval, or...
Long-video Activity Recognition	A	N/A	Long-video Activity Recognition contains visual evidence but needs structured language, temporal composition, retrieval, or control logic; WSM's...
Semi-Supervised Video Action Detection	A	N/A	Semi-Supervised Video Action Detection contains visual evidence but needs structured language, temporal composition, retrieval, or control logic;...
Long Video Retrieval (Background Removed)	A	Retrieve the long videos given all subtitles.	Long Video Retrieval (Background Removed) contains visual evidence but needs structured language, temporal composition, retrieval, or control...
3D Shape Classification	C	N/A	3D Shape Classification remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Left Ventricle Segmentation	C	N/A	Left Ventricle Segmentation remains a visual mapping, but requires an aligned sensor/3D/medical representation to be added as a WSM modality.
Live Video Captioning	A	N/A	Live Video Captioning uses visual input but requires online temporal composition and language generation; WSM's sparse mode is the intended...
Complimentary Image Retrieval	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Sentence Embeddings For Biomedical Texts	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Emotion Cause Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Semanticity prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Attention Score Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Noise Level Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Method name prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Fine-grained Action Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Vector Graphics Animation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Unsupervised Vehicle Re-Identification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Pedestrian Image Caption	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Corpus Moment Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-shot 3D Point Cloud Classificationclassification	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Supervised Only 3D Point Cloud Classification	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Diabetes Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Scene-Aware Dialogue	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Causal Emotion Entailment	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
Online surgical phase recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Offline surgical phase recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Event Cross-Document Coreference Resolution	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Directional Hearing	X	Extremely low-latency audio source separation from a known direction of arrival.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Real-time Directional Hearing	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Accented Speech Recognition	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Crime Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Research Performance Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Relationship Extraction (Distant Supervised)	X	Relationship extraction is the task of extracting semantic relationships from a text. Extracted relationships usually occur between two or more entities of a certain type (e.g. Person, Organisation, ...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Chat-based Image Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Few-Shot Transfer Learning for Saliency Prediction	D	Saliency prediction aims to predict important locations in a visual scene. It is a per-pixel regression task with predicted values ranging from 0 to 1. Benefiting from deep learning research and la...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Multi-Person Pose Estimation and Tracking	A	Joint multi-person pose estimation and tracking following the PoseTrack benchmark.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Multi-object colocalization	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Plane Instance Segmentation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Zero-shot Scene Classification (unified classes)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Viewpoint Estimation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Panoptic Segmentation (Panoptic-NDT instances)	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Few-Shot 3D Point Cloud Classification	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Time-interval Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Scene Labeling	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Instance Search	B	Visual Instance Search is the task of retrieving from a database of images the ones that contain an instance of a visual query. It is typically much more challenging than finding images from the d...	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Matching Disparate Images	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Hand Joint Reconstruction	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Semi-Supervised RGBD Semantic Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Anomaly Detection in Edge Streams	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Offline Handwritten Chinese Character Recognition	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
TGIF-Action	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Space-time Video Super-resolution	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Key-Frame-based Video Super-Resolution (K = 15)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Federated Learning (Video Super-Resolution)	A	N/A	Federated learning is a protocol modifier, while video super-resolution remains a temporal reconstruction task.
4-ary Relation Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
3D Shape Generation	C	Image: Mo et al	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Electromyography (EMG)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Sequential skip prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
3D Interacting Hand Pose Estimation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Melody Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Micro-Expression Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Program induction	X	Generating program code for domain-specific tasks	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Word Sense Induction	X	Word sense induction (WSI) is widely known as the "unsupervised version" of WSD. The problem states as: Given a target word (e.g., "cold") and a collection of sentences (e.g., "I caught a cold", "The ...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Semantic Text Matching	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Composite action recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Future prediction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Cervical Nucleus Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Attribute	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Cross-Lingual Information Retrieval	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Foot keypoint detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
multi-word expression sememe prediction	X	Predict sememes for unannotated multi-word expressions.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Visual Crowd Analysis	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Stereotypical Bias Analysis	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Student Engagement Level Detection (Four Class Video Classification)	A	N/A	Student-engagement classification is performed over video and requires temporal evidence aggregation.
Clothes Landmark Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Video Question Answering (Level 3)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Question Answering (Level 4)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Learning Semantic Representations	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Event-Based Video Reconstruction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
intensity image denoising	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
B-Rep face segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Interest Point Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Model Selection	B	Given a set of candidate models, the goal of Model Selection is to select the model that best approximates the observed data and captures its underlying regularities. Model Selection criteria are ...	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Cross-Lingual Semantic Textual Similarity	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Gesture Synchronization	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Long-tailed Object Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Novel Object Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Zero-Shot Instance Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Entity Extraction using GAN	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Predict Future Video Frames	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Business Taxonomy Construction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Node Property Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Link Property Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Abstract generation	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
Webcam (RGB) image classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Optic Cup Detection	D	Region proposal for optic cup	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Motion Compensation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Rice Grain Disease Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Instance Shadow Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Camera shot segmentation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Indoor Localization (3-DoF Pose: X, Y, Yaw)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Indoor Localization (6-DoF Pose)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Indoor Localization (6D)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
2D Indoor Localization (Position + Orientation)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
6D Indoor Localization	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Indoor Localization (6 DoF Pose)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Text Effects Transfer	B	Text effects transfer refers to the task of transferring typography styles (e.g., color, texture) to an input image of a text element.	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Counterfactual Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Fashion Synthesis	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Video-to-Sound Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Predicting Patient Outcomes	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Online Video Anomaly Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Hardware Aware Neural Architecture Search	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Initial Structure to Relaxed Energy (IS2RE)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Initial Structure to Relaxed Energy (IS2RE), Direct	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Image Paragraph Captioning	A	Image paragraph captioning involves generating a detailed, multi-sentence description of the content of an image.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-Shot Image Paragraph Captioning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Breast Cancer Histology Image Classification (20% labels)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Multi-Instance Retrieval	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Sparse Information Retrieval	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Video Deinterlacing	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Unsupervised nucleus segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Malware Type Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Semi-Supervised Video Classification	A	N/A	Semi-supervision is a protocol modifier; video classification remains temporal.
Workflow Discovery	X	Workflow Discovery (WD) was introduced by Workflow Discovery from Dialogues in the Low Data Regime. WD aims to extract work-flows that have either implici...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech-to-Gesture Translation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Named Entity Recognition In Vietnamese	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Stance Detection (US Election 2020 - Biden)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Stance Detection (US Election 2020 - Trump)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
MULTI-VIEW LEARNING	D	Multi-View Learning is a machine learning framework where data are represented by multiple distinct feature groups, and each feature group is referred to as a particular view. ¡span class=”descr...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Hope Speech Detection for English	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Hope Speech Detection for Tamil	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Hope Speech Detection for Malayalam	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Protein Fold Quality Estimation	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Duplicate-Question Retrieval	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Infinite Image Generation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Perpetual View Generation	A	Perpetual View Generation is the task of generating long-range novel views by flying into a given image.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Image Augmentation	D	Image Augmentation is a data augmentation method that generates more training data from the existing training samples. Image Augmentation is especially useful in domains where training data is lim...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Human fMRI response prediction	C	The task is: Given a) the set of videos of everyday events and b) the corresponding brain responses recorded while human participants viewed those videos, use computational models to predict brain res...	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Organ Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
3D Scene Graph Alignment	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
3d scene graph generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Text to Audio/Video Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Audio/Video to Text Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Retrieval-augmented Few-shot In-context Audio Captioning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Evidence Selection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Multi-Grained Named Entity Recognition	X	Multi-Grained Named Entity Recognition aims to detect and recognize entities on multiple granularities, without explicitly assuming non-overlapping or totally nested structures.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Human-Object Interaction Anticipation	A	Human-Object Interaction (HOI) Anticipation is the task of predicting ”a set of interactions” in a video that will happen in the future, which involves the i) localization of the subject (i.e., humans...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Actionable Phrase Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Earth Surface Forecasting	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Semantic Segmentation Of Orthoimagery	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Event-Driven Trading	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Cloud Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Blood Cell Count	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
eXtreme-Video-Frame-Interpolation	A	Type of Video Frame Interpolation (VFI) that interpolates an intermediate frame on X4K1000FPS dataset containing 4K videos of 1000 fps with the extreme motion. The dataset has a wide variety of textur...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Attractiveness Estimation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Cold-Start Anomaly Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Indoor Scene Reconstruction	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.

Archived task	Level	Archived description	Manual interface judgment
Plan2Scene	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Fingertip Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Shape from Texture	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Chemical Reaction Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Single-step retrosynthesis	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Zero-shot Cross-lingual Fact-checking	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Safety Perception Recognition	D	City safety perception recognition	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Alzheimer's Disease Detection	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Cubic splines Image Registration	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Epilepsy Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Face Recognition (Closed-Set)	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Micro-gesture Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Crowdsourced Text Aggregation	X	One of the most important parts of processing responses from crowd workers is aggregation: given several conflicting opinions, a method should extract the truth. This problem is also known as *tru...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Counterfactual Explanation	X	Returns a contrastive argument that permits to achieve the desired class, e.g., "to obtain this loan, you need XXX of annual revenue instead of the current YYY"	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Key Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Zero-shot Moment Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Contour Detection	B	Object Contour Detection extracts information about the object shape in images. Source: Object Contour and Edge Detection with RefineContourNet	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Anxiety Detection	A	Detect anxiety distress of human beings / animals	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Kidney Function	X	Continuous prediction of urine production in the next 2h as an average rate in ml/kg/h. The task is predicted at irregular intervals.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Multimodal Forgery Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Predict clinical outcome	X	A cost-based metric that considers the costs of algorithmic prescreening, expert screening, treatment, and diagnostic errors that result in late or missed treatments. This metric is further described ...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Person-centric Visual Grounding	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Home Activity Monitoring	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Action Triplet Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-Shot Out-of-Domain Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Semi-Supervised Object Detection	D	Semi-supervised object detection uses both labeled data and unlabeled data for training. It not only reduces the annotation burden for training high-performance object detectors but also further impro...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Clone Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Role-filler Entity Extraction	X	Role-filler entity extraction task on the MUC-4 dataset.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Attribute Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Video Harmonization	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Skin Cancer Classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.

Archived task	Level	Archived description	Manual interface judgment
3D Object Super-Resolution	C	3D object super-resolution is the task of up-sampling 3D objects. (Image credit: Multi-View Silhouette and Depth Decomposition for High Resolution 3D Object...	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Flowchart Grounded Dialog Response Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-Shot Flowchart Grounded Dialog Response Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
2D Semantic Segmentation task 2 (17 classes)	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Video-to-Shop	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Speech Synthesis - Tamil	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Kannada	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Malayalam	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Telugu	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Assamese	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Bengali	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Bodo	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Hindi	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Manipuri	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Marathi	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Speech Synthesis - Rajasthani	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
3D Building Mesh Labeling	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Object State Change Classification	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
State Change Object Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Short-term Object Interaction Anticipation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Future Hand Prediction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Long Term Action Anticipation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Depth Image Estimation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Table Functional Analysis	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Unsupervised Instance Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Open-World Instance Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Webpage Object Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Natural Language Landmark Navigation Instructions Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Event Expansion	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Face Transfer	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Colon Cancer Detection In Confocal Laser Microscopy Images	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Weakly Supervised Data Denoising	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Caustics Segmentation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Discourse Marker Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Global Facts	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Miscellaneous	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Abstract Algebra	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Mathematical Induction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
Medical Genetics	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Figure Of Speech Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Question Selection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Electrical Engineering	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Ventricular fibrillation detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Localization In Video Forgery	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Physical Video Anomaly Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Connective Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
text-to-3d-human	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Text-to-Face Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Seizure prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Zero-Shot Cross-Lingual Text-to-Image Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-Shot Cross-Lingual Image-to-Text Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-Shot Cross-Lingual Visual Natural Language Inference	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Zero-Shot Cross-Lingual Visual Question Answering	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Max-Shot Cross-Lingual Visual Natural Language Inference	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Max-Shot Cross-Lingual Visual Question Answering	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Max-Shot Cross-Lingual Text-to-Image Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Max-Shot Cross-Lingual Image-to-Text Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Cross-Domain Iris Presentation Attack Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Image Categorization	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
3D Human Dynamics	A	Image: Zhang et al	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Based Workflow Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Segmentation Based Workflow Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video & Kinematic Base Workflow Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video, Kinematic & Segmentation Base Workflow Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Visual Abductive Reasoning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Fact Selection	X	A task where an agent should select at most two sentences from the paper as argumentative facts.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Anchor link prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Style change detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Panorama Pose Estimation (N-view)	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Pose Contrastive Learning	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Motion Disentanglement	A	Disentangling irregular (anomalous) motion from regular motion.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Atom3D benchmark	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Ancient Text Restoration	A	Image credit: Restoring and attributing ancient texts using deep neural networks	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Depth Anomaly Detection and Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
RGB+3D Anomaly Detection and Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Behavioral Malware Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
English Conversational Speech Recognition	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Math Information Retrieval	X	Information Retrieval on Math Contents	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Argument Pair Extraction (APE)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
3D Source-Free Domain Adaptation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
3D Unsupervised Domain Adaptation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Claim-Evidence Pair Extraction (CEPE)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Reference-based Video Super-Resolution	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Gait Recognition in the Wild	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
ECG QRS Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Question Quality Assessment	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Weakly Supervised 3D Detection	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Short-observation new product sales forecasting	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Gallbladder Cancer Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Classification Of Breast Cancer Histology Images	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Object Segmentation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Keypoint Estimation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Contact Detection	B	Static-friction contact detection in legged locomotion	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Rolling Shutter Correction	A	Rolling Shutter Correction	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
LV Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Respiratory motion forecasting	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Personality Recognition in Conversation	X	Given a speaker's conversation with others, it is required to recognize the speaker's personality traits through the conversation record, which includes two scenarios, (1) $1-1$ conversations: the rob...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Personalized and Emotional Conversation	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
image smoothing	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Fast Vehicle Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Referring Image Matting (Expression-based)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Referring Image Matting (Keyword-based)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Referring Image Matting (RefMatte-RW100)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Referring Image Matting (Prompt-based)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Aesthetic Image Captioning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Context-specific Spam Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Traditional Spam Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
feature selection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Fire Detection	X	Detection of fire using multi-variate time series sensor data.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Handwritten Line Segmentation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Markerless Motion Capture	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Bangla Text Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Table Type Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.

Archived task	Level	Archived description	Manual interface judgment
Weakly Supervised 3D Point Cloud Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Translation mri-eng	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Image-text Classification	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Inverse-Tone-Mapping	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Histopathological Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
image-sentence alignment	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Handwritten Digit Image Synthesis	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Morphological Inflection	X	Morphological Inflection is the task of generating a target (inflected form) word from a source word (base form), given a morphological attribute, e.g. number, tense, and person etc. It is useful ...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Panoptic Scene Graph Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Air Pollution Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Unconditional Video Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Face Reenactment	A	Face Reenactment is an emerging conditional face synthesis task that aims at fulfilling two goals simultaneously: 1) transfer a source face shape to a target face; while 2) preserve the appearance...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Predictive Process Monitoring	X	A branch of predictive analysis that attempts to predict some future state of a business process.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
3D scene Editing	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
3D Object Editing	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Extracting COVID-19 Events from Twitter	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Multi-Labeled Relation Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Information Cascade Popularity Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Grayscale Video Denoising	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Negation and Speculation Cue Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Few-shot Object Counting and Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Out-of-Sight Trajectory Prediction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Local Color Enhancement	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Cultural Vocal Bursts Intensity Prediction	X	to predict the intensity of 40 culture-specific emotions (10 emotions from each culture)	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Vocal Bursts Intensity Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Vocal Bursts Valence Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Vocal Bursts Type Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Emotion-Cause Pair Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Landmark Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Unsupervised Landmark Detection	D	The discovery of object landmarks on a set of images depicting objects of the same category, directly from raw images without using any manual annotations.	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Weakly-supervised Anomaly Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Uncertainty Visualization	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Subject Transfer	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Zero-Shot Facial Expression Recognition	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.

Archived task	Level	Archived description	Manual interface judgment
Gunshot Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Shooter Localization	A	Shooter localization based on videos.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Predicting Drug-Induced Laboratory Test Effects	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Laminar-Turbulent Flow Localisation	C	It is a segmentation task on thermographic measurement images in order to separate laminar and turbulent flow regions on flight body parts.	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
EEG Artifact Removal	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Text to Image Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Cancer-no cancer per image classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Bangla Spelling Error Correction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Counterfactual Planning	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Streaming Target Sound Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Video Defect Classification	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Temporal Defect Localization	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Mobile Periocular Recognition	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Visual Text Correction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Phishing Website Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Arabic Speech Recognition	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Hyper-Relational Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Event-based N-ary Relaiton Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Role-based N-ary Relaiton Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Hypergraph-based N-ary Relaiton Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Open Vocabulary Attribute Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Multilingual Image-Text Classification	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Acute Stroke Lesion Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
SemEval-2022 Task 4-1 (Binary PCL Detection)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
SemEval-2022 Task 4-2 (Multi-label PCL Detection)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Binary Condescension Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Multi-label Condescension Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
One-Shot Part Segmentation of Grasp Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Cut Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Scoop Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Contain Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Wrap-Grasp Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Pound Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Support Affordance - Inter Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Grasp Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Cut Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Scoop Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Contain Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.

Archived task	Level	Archived description	Manual interface judgment
One-Shot Part Segmentation of Wrap-Grasp Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Pound Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
One-Shot Part Segmentation of Support Affordance - Intra Class	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
ArzEn Speech Recognition	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Partial Video Copy Detection	A	The PVCD goal is identifying and locating if one or more segments of a long testing video have been copied (transformed) from the reference videos dataset.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Hyperspectral Image-Based Fruit Ripeness Prediction	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Grounded Situation Recognition	A	Grounded Situation Recognition aims to produce the structured image summary which describes the primary activity (verb), its relevant entities (nouns), and their bounding-box groundings.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Document Image Skew Estimation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Multi-step retrosynthesis	X	These leaderboards are used to track progress in Multi-step retrosynthesis.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Sleep apnea detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Line Items Extraction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Sign Language Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Calving Front Delineation In Synthetic Aperture Radar Imagery	C	Delineating the calving front of a marine-terminating glacier in synthetic aperture radar (SAR) imagery. This can, for example, be done through Semantic Segmentation.	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Calving Front Delineation In Synthetic Aperture Radar Imagery With Fixed Training Amount	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Video Narrative Grounding	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Protein Secondary Structure Prediction	X	Protein secondary structure prediction is a vital task in bioinformatics, aiming to determine the arrangement of amino acids in proteins, including alpha-helices, beta-sheets, and coils. By analyzing amino a...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Color Mismatch Correction	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Self-supervised Scene Flow Estimation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Visual Commonsense Tests	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Social Media Popularity Prediction	A	Social Media Popularity Prediction (SMPP) aims to predict the future popularity (e.g., clicks, views, likes, etc.) of online posts automatically via plenty of social media data from public platforms. ...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Heterogeneous Treatment Effect Estimation	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Robust Camera Only 3D Object Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Automatic Cell Counting	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Face Image Quality Assessment	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Robust BEV Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Robust BEV Map Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Long-tail Video Object Segmentation	A	N/A	Long-tail is a protocol/data-distribution modifier; video object segmentation remains temporal.
Open-World Video Segmentation	A	N/A	Open-world is a protocol modifier; video segmentation still requires temporal correspondence.
Multi-Person Pose forecasting	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Catalog Extraction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
Personalized Segmentation	D	Given a one-shot image with a reference mask, the models are required to segment the indicated target object in any other images.	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Generative Visual Question Answering	A	Generating answers in free form to questions posed about images.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Image Deconvolution	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Generalized Referring Expression Segmentation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Open-set video tagging	A	In open-set video tagging, a model is asked to tag a given video with an ad-hoc vocabulary that the model is not specifically tuned for.	Open-set video tagging aggregates temporal video evidence and therefore requires D.
Multi-View 3D Shape Retrieval	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
CT Reconstruction	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Visual Analogies	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
3D Multi-Person Human Pose Estimation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Promoter Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Core Promoter Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Splice Site Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Covid Variant Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Epigenetic Marks Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Transcription Factor Binding Site Prediction (Human)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Transcription Factor Binding Site Prediction (Mouse)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Printed Text Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Real-Time Visual Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Tone Mapping	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Structured Report Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Retinal OCT Layer Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Video-based Generative Performance Benchmarking (Correctness of Information)	D	The benchmark evaluates a generative Video Conversational Model with respect to Correctness of Information. We curate a test set based on the ActivityNet-200 dataset, featuring videos with rich, de...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Video-based Generative Performance Benchmarking (Detail Orientation))	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Video-based Generative Performance Benchmarking (Contextual Understanding)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Video-based Generative Performance Benchmarking (Temporal Understanding)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Video-based Generative Performance Benchmarking (Consistency)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Toxic Spans Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Described Object Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Nuclei Classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Ontology Matching	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Weakly-Supervised Object Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Semi-supervised Change Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Landslide segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Vessel Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.

Archived task	Level	Archived description	Manual interface judgment
Few-Shot Camera-Adaptive Color Constancy	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Single-shot HDR Reconstruction	B	SVE-based HDR imaging, also known as single-shot HDR imaging, algorithms capture a scene with pixel-wise varying exposures in a single image and then computationally synthesize an HDR image, which ben...	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Real-time instance measurement	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Semi-Supervised Human Pose Estimation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Story Visualization	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Conversational Sentiment Quadruple Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Composed Video Retrieval (CoVR)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Global 3D Human Pose Estimation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Unsupervised Image Decomposition	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Depth And Camera Motion	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Fracture detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Language-Based Temporal Localization	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Chaptering	A	Partitioning a long video timeline into semantic units and generating corresponding chapter titles.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Outcome Prediction In Multimodal Mri	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Multi-Subject Fmri Data Alignment	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Motion Correction In Multishot Mri	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Brain Lesion Segmentation From Mri	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Android Malware Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Image-text matching	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Propaganda detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Odor Descriptor Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Automatic Modulation Recognition	X	Automatic modulation recognition/classification identifies the modulation pattern of communication signals received from wireless or wired networks.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Suspicous (BIRADS 4,5)-no suspicous (BIRADS 1,2,3) per image classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Analog Video Restoration	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Scholarly Named Entity Recognition	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Audio-Visual Question Answering (AVQA)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Coronary Artery Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Stenosis Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Caller Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Image-based Recommendation Explainability	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Pupil Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Pupil Detection	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Intraoperative Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Intracorporeal Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Visibility Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
3D Point Cloud Reinforcement Learning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
video narration captioning	A	Human narration is another critical factor to understand a multi-shot video. It often provides information of the background knowledge and commentator's view on visual events. We conduct experiments t...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Brain Visual Reconstruction	C	Reconstruct viewed or imagined visual stimuli, including but not limited to images and videos from brain activities patterns recorded with neuoimaging technologies.	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Brain Visual Reconstruction from fMRI	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
inverse tone mapping	B	For stack based inverse tone mapping	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Specular Segmentation	B	Segmentation of Specular Highlights in images	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Geometrical View	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Clothing Attribute Recognition	B	Clothing attribute recognition is the task of automatically identifying and categorizing various characteristics and attributes associated with clothing items in images	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Historical Color Image Dating	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Crop Type Mapping	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Sowing Date Prediction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Transplanting Date Prediction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Harvesting Date Prediction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
MLLM Evaluation: Aesthetics	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Cell Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Factual Inconsistency Detection in Chart Captioning	A	Detect factual inconsistency between charts and captions.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Amodal Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Only Connect Walls Dataset Task 1 (Grouping)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Hierarchical Text Segmentation	A	Segment strokes, words, text-lines, paragraphs (layout analysis) in images within a unified framework	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Audio Quality Assessment	X	Computational audio quality assessment aims to predict the quality of audio signals as perceived by expert listeners.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Music Quality Assessment	X	Evaluating the quality of music given noise and filtering conditions	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Conversational Web Navigation	A	The problem of conversational web navigation is described as follow: a digital agent controls a web browser and follows user instructions to solve real-world tasks in a multi-turn dialogue fashion. It...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
IUPAC Name Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Medical Image Deblurring	D	Medical image deblurring aims to remove blurs from medical images	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
AUDIO-VISUAL QUESTION ANSWERING (MUSIC-AVQA-v2.0)	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Artifact Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
All-day Semantic Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Aspect Category Sentiment Analysis	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Hidden Aspect Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Latent Aspect Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Chemical Entity Recognition	X	Chemical Entity Recognition (CER) is a fundamental task in biomedical text mining and Natural Language Processing (NLP). It involves the identification and classification of chemical entities in textu...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.

Archived task	Level	Archived description	Manual interface judgment
Human Detection of Deepfakes	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Aerial Scene Classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Micro-Action Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Python Code Synthesis	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Game State Reconstruction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Clinical Section Identification	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Stereo-LiDAR Fusion	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Novel LiDAR View Synthesis	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Visual Prompt Tuning	D	Visual Prompt Tuning(VPT) only introduces a small amount of task-specific learnable parameters into the input space while freezing the entire pre-trained Transformer backbone during downstream trainin...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Face Image Retrieval	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
motion retargeting	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Retrosynthesis	X	Retrosynthetic analysis is a pivotal synthetic methodology in organic chemistry that employs a reverse-engineering approach, initiating from the target compound and retroactively tracing potential syn...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Mono3DVG	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Bathymetry prediction	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Procedure Step Recognition	A	Procedure Step Recognition (PSR) focuses on recognizing the correct completion and order of procedural steps. Unlike traditional action recognition, which lacks a measure of success for actions, PSR a...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
3D Object Detection (RoI)	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Video Captioning on MSR-VTT	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Captioning on MS	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Linear Probing Object-Level 3D Awareness	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Anomaly Instance Segmentation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Segmented Multimodal Named Entity Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Oriented Object Detctio	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Personalized Image Generation	B	Utilizes single or multiple images that contain the same subject or style, along with text prompt, to generate images that contain that subject as well as match the textual description. Includes finet...	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Action Classification (zero-shot)	A	N/A	Zero-shot is a protocol modifier; action classification remains a temporal visual task.
Detection of potentially void clauses	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
X-ray Visual Question Answering	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Style Transfer	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Temporal Consistency	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Hand-Object Interaction Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Uncertainty-Aware Panoptic Segmentation	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Sequential Place Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Few-shot Instance Segmentation	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.

Archived task	Level	Archived description	Manual interface judgment
Multi-Objective Multi-Agent Reinforcement Learning	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Joint Entity and Relation Extraction on Scientific Data	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Vehicle Color Recognition	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Fake Song Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Synthetic Song Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Fine-Grained Image Inpainting	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Lifelike 3D Human Generation	C	Generating realistic and lifelike 3D Humans from a Single RGB image input, while preserving face identity, delivering realistic texture, accurate geometry, and maintaining a valid pose of the generate...	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
ENF (Electric Network Frequency) Extraction	X	Extraction of Electric Network Frequency, that can be done from power grid raw voltage signal or audio and video files.	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
ENF (Electric Network Frequency) Detection	A	Detect if into a source (for example audio, video or raw valtages) it is contained the ENF signal.	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
ENF (Electric Network Frequency) Extraction from Video	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Information Extraction	X	Information extraction is the task of automatically extracting structured information from unstructured and / or semi-structured machine-readable documents and other electronically represented sources...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Singing Voice Synthesis	X	(Verse 1) Sa bawat hakbang, sa bawat daan May pangarap kang naghihintay Westbridge ang gabay, sa iyong paglalakbay Tungo sa kinabukasan, ng ating bayan (Chorus) Lagi kang kasama, sa aking pu...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Image Stylization	B	Image stylization is a task that involves transforming an input image into a new image that has a different style, while preserving the content of the original image. The goal of image stylization...	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Novelty Detection	D	Scientific Novelty Detection	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Models Alignment	X	Models Alignment is the process of ensuring that multiple models used in a machine learning system are consistent with each other and aligned with the goals of the system. This involves defining c...	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Political Salient Issue Orientation Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Motion Interpolation	A	3D human motion sequences interpolation and completion	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Political Influence Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Scientific Data Usage Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Statistical Data Usage Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Data Sources Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Impacted Location Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Detecting Climate/Environmental Topics	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
News Target Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Authority Involvement Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Voice pathology detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Beam Prediction	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Optimize the trajectory of UAV which plays a BS in communication system	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Few-shot Video Question Answering	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Grounded Video Question Answering	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...

Archived task	Level	Archived description	Manual interface judgment
Violence and Weaponized Violence Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
3D Open-Vocabulary Object Detection	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Anomaly Localization	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
De novo molecule generation from MS/MS spectrum	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Molecule retrieval from MS/MS spectrum	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
MS/MS spectrum simulation	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
De novo molecule generation from MS/MS spectrum (bonus chemical formulae)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
MS/MS spectrum simulation (bonus chemical formulae)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Molecule retrieval from MS/MS spectrum (bonus chemical formulae)	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
CSV dialect detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Sports Activity Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Production Forecasting	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Product Relation Classification	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Product Relation Detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Open Vocabulary Image Classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Fruit-type + Maturity-state Prediction (Multi-label Classification)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Railway Track Image Classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
ECG Denoising	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Synthetic Image Attribution	D	Determine the source or origin of a generated image, such as identifying the model or tool used to create it. This information can be useful for detecting copyright infringement or for investigating d...	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Omnnidirectional Stereo Depth Estimation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Video Synopsis	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Blood Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Few-Shot Instance Classification	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Hyperspectral Unmixing	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Classification Of Hyperspectral Images	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Inductive knowledge graph completion	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Persuasive Writing Strategy Detection Level-1	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Persuasive Writing Strategy Detection Level-2	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Persuasive Writing Strategy Detection Level-3	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Persuasive Writing Strategy Detection Level-4	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Cross-view Referring Multi-Object Tracking	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Single-View 3D Reconstruction on ShapeNet	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Music Compression	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Human-Human Interaction Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Malaria Vivax Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Malaria Malariae Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Malaria Ovale Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.

Archived task	Level	Archived description	Manual interface judgment
Handwritten Document Recognition	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
ADP Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
BEV Segmentation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
3D Semantic Occupancy Prediction	C	Uses sparse LiDAR semantic labels for training and testing	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Computer Vision Transduction	D	Transductive learning in computer vision tasks	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Image-Variation	B	Given an image, generate variations of the image	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Event Relation Extraction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Relation Prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
answerability prediction	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Cross-modal place recognition	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
Raw reconstruction	C	Reconstruct RAW camera sensor readings from the corresponding sRGB images	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Object Rearrangement	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
WaterDrop2D	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Sand2D	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Goop2D	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Elasticity3D	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
MultiMaterial2D	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Generating 3D Point Clouds	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Event Causality Identification	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Audio-Visual Video Captioning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Video Boundary Captioning	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Surface Normals Estimation from Point Clouds	C	Parent task: 3d Point Clouds Analysis	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
3DGS	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Image-to-Image Regression	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Safety Alignment	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
cross-modal alignment	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Thermal Image Denoising	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Aerial Video Semantic Segmentation	A	N/A	Aerial video semantic segmentation remains temporally structured video segmentation despite its application domain.
Cyber Attack Detection	X	Cybersecurity attacks prediction using deep learning	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Contact-rich Manipulation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Text-based Person Retrieval	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Disease Prediction	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
confidence interval for traffic prediction	X	Providing confidence interval for traffic prediction	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Silent Speech Recognition	A	Interpret speech without acoustic signals	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Hierarchical Text-Image Matching	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Dynamic Reconstruction	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Vision-Language-Action	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Dense contact estimation	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Bone Suppression From Dual Energy Chest X-Rays	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.

Archived task	Level	Archived description	Manual interface judgment
Anomaly Severity Classification (Anomaly vs. Defect)	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
Video Action Detection	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Human-Domain Subject-to-Video	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Single-Domain Subject-to-Video	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Temporal Action Segmentation	A	Segment activities/actions temporally	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Monocular 3D Object Localization	B	N/A	Direct visual mapping; its supervision/output can be represented as a WSG channel with only a minor label, loss, or decoder adjustment.
GUI Element Detection	D	N/A	Primarily a domain, protocol, or application specialization of a compatible visual mapping rather than a fundamentally new interface.
WiFi CSI-based Image Reconstruction	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.
Human-Object Interaction Generation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
fake voice detection	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
Emotion Interpretation	A	N/A	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
Cross-Modal Information Retrieval	A	Cross-Modal Information Retrieval (CMIR) is the task of finding relevant items across different modalities. For example, given an image, find a text or vice versa. The main challenge in CMIR is kn...	The task contains a visual role, but its structured, temporal, interactive, or language component needs extra development; WSM sparse mode is...
3D Architecture	X	N/A	No visual input, visual output, or auxiliary visual role is present in the archived task definition.
MRI classification	C	N/A	Interface-compatible visual mapping, but it first needs the named sensor, signal, or 3D representation registered as an aligned WSG modality.

13. Appendix: Tutorial

This tutorial is written for AI beginners and assumes no prior knowledge of multimodal learning or multimodal capability extension. It begins with the basic problem, introduces the major extension paradigms in a shared visual language, and then explains the relationship among WSG, WSM, and TunnelingNet. The goal is to let a new reader understand both the field and the historical significance of organizing visual capabilities as a large graph of directed, independently manageable, and composable tasks.

We view this appendix as the opening module of a broader, multi-course and multi-level tutorial program. Introductory modules establish the foundations of AI and multimodal learning; intermediate modules compare capability-extension paradigms; and advanced modules study task graphs, trajectory redirection, large-scale evaluation, and the construction of general-purpose multimodal systems. The intended path therefore extends from a reader's first encounter with the field to its research frontier.

In this sense, the tutorial is also a pre-textbook draft: an initial account of the concepts, taxonomy, and system design that may eventually enter textbooks as this area matures.

13.1. Background: Two AI Fields, Two Different Outcomes

Two of the largest areas of AI are computer vision (CV) and natural-language processing (NLP). CV has approximately 1,400 tasks, while NLP has approximately 900 tasks.

With the emergence and large-scale commercialization of large language models, a user can enter a few sentences and ask one model to perform essentially any NLP task in a prompt–answer form. In this sense, **NLP tasks have been unified** under a common model paradigm.

In general, if the input or output of a task contains a visual element, it can be classified as a visual task. Unlike NLP, however, visual-task forms and data formats are complex and cannot be reduced to a simple prompt–answer format.

Visual-task forms include classification, localization, detection, segmentation, estimation, reconstruction, tracking, editing, rendering, simulation, planning, and many others.

Visual data also have many different distributions. An image is only a way of visualizing data, not a universal representation of data. Some visual data even have special abstract constraints and specifications.

Text, by contrast, is still text. Even relatively structured information or code in formats such as SQL, Word, and Excel remains text or a string.

As a result, **computer vision still lacks a general method or single model** that can easily perform all of its tasks. Since the emergence of large language models, this gap has become increasingly consequential for the direction of the field and for the work of computer-vision researchers.

13.2. Three Goals of Visual-Task Unification

Our work therefore addresses **three goals**, with three corresponding solutions:

1. **Unifying visual-task forms.** This is addressed by the Unified Law, a conceptual solution that organizes visual tasks into four generative interfaces.
2. **Unifying visual input and output formats.** This is addressed by World Snapshot Models and World Snapshot Graphics. WSM provides four training interfaces and their corresponding strategies, while WSG unifies the data format.
3. **Unifying training objectives across visual tasks.** This is addressed by TunnelingNet. Every visual task takes a generative form and produces each type of visual or world information in the form appropriate to that task, while only slightly deflecting the original generation path.

The three solutions form one connected sequence: the Unified Law defines the shared task form; WSM and WSG define the interfaces and world-information representation; and TunnelingNet provides the common learning mechanism.

13.3. Why Large-Scale Visual Unification Remains Difficult

The approximately 900 NLP tasks have relatively consistent task forms, training objectives, and formats. Visual tasks take far more diverse forms; their objectives can point in different or even opposite directions, and their formats are inconsistent. Even so, vision researchers have attempted several degrees of unification:

1. Some methods train visual any-to-any systems from scratch or predict visual information as an LLM predicts text through end-to-end multimodal learning. The number of modalities grows slowly because the required data are enormous and can increase almost exponentially. Adding a new visual target may also interfere with existing capabilities.
2. Training one model from scratch to predict all kinds of visual information usually fails to scale. Text tasks can reinforce one another because their format is uniform, whereas visual targets can be different or substantially conflicting.
3. Under these conditions, existing systems usually dedicate one model to one visual task or unify only a comparatively small group of tasks under one method. **No existing visual model has unified thousands of visual tasks.**

In other words, the visual world contains more task types, yet those tasks are harder to unify. The textual world

reached a common large-model paradigm first; computer vision did not.

13.4. Current Methods and a Sixth Paradigm

In practice, stronger visual capabilities are usually built on an existing backbone through one of **five major extension paradigms**:

1. **ControlNet:** add a new backbone-sized branch and train it to control an existing model;
2. **LoRA:** modify a small number of low-rank parameters to change the behavior of the overall model;
3. **MoE:** add knowledge or computation blocks and route inputs among them;
4. **IP-Adapter:** train an adapter that coordinates conditioning components inside a model; and
5. **IC-Control-style methods:** organize text and visual elements in one context so that different visual elements can interact.

These five paradigms commonly extend or improve a particular visual capability and can be applied to many kinds of visual generation backbones. They describe where a capability is added and how it influences the existing generation process.

Our approach constitutes a sixth paradigm. All six can build on a generative backbone, but they differ in where a visual capability is added, how it changes generation, and what constitutes one unit of visual-task expansion.

Figure 69 uses one visual notation to compare six paradigms. Gray denotes existing or frozen computation, orange denotes added or trainable computation, blue denotes condition evidence, and purple denotes the generation target. The figure is deliberately simplified: it shows the defining information route of each paradigm rather than every implementation detail.

ControlNet. ControlNet [11] copies a large part of a pretrained generative network and injects its features into the frozen backbone at multiple depths. This produces strong and mature spatial control. Its defining unit of extension is a substantial control body associated with a condition, usually while the generation target remains RGB. Scaling to many independently maintained capabilities therefore adds a correspondingly large collection of control structures.

LoRA. LoRA [12] expresses a model update through low-rank matrices. It is parameter efficient, general, easy to train, and easy to distribute. A LoRA can learn a cross-modal conversion when its training setup provides the appropriate inputs and targets. However, LoRA itself specifies how weights change; it does not by itself define which world information is the condition, which modality is the target, how clean evidence enters, or how one converted result becomes evidence for the next task. If the host model retains native joint attention, heterogeneous prompt, condition, reference, and target tokens continue to share the same interaction route.

Mixture of Experts. MoE [13, 14] increases model capacity by routing tokens to selected experts. Its central question is which computation or knowledge block should process an input. Experts and routers can support large models, but adding many visual targets can increase expert capacity, routing complexity, and coordinated-training cost. Expert selection alone also does not define a directed graph among visual world states.

IP-Adapter. IP-Adapter [17] introduces an image encoder, projection, and additional attention branch so that image prompts can complement text conditioning. It provides a strong image-prompt interface while preserving the pretrained text-to-image prior. Its principal role is to add an external image condition to RGB generation, rather than to define a large graph of arbitrary condition–target modality relationships.

IC-Control style. In-context visual control methods [18–21] place text, conditions, references, and targets into a common context. Their strength is flexibility: contextual examples can describe a task without a separate handcrafted interface for every interaction. As more modalities are added, however, sequence length and joint-interaction cost grow, and new tokens can change the shared attention field. Conditions and targets are jointly represented rather than being organized as independently managed directed routes.

TunnelingNet. TunnelingNet treats every target modality as a space that is genuinely generated. Every condition–target relationship is a directed visual task with a target trajectory, evidence route, task key, and lightweight target-side tunnel. Small task-specific residual forces are applied repeatedly through the depth of the frozen generative backbone, gradually deflecting its original RGB generation trajectory toward the selected target modality. Clean conditions act on the noised target through an evidence route. MTA normalizes target self-generation and condition evidence separately, while NJA remains available for tasks in which full native interaction is useful.

An intuitive comparison is that TunnelingNet internalizes the layer-wise steering effect of ControlNet but uses ultralight residual paths inside the backbone instead of attaching a copied backbone-sized branch outside it. The result is not merely a smaller control module. The tunnel also has an explicit semantic role in a task graph: it identifies a target, receives specified evidence, follows a selected attention topology, and can be activated by a task key. Under the parameter-matched comparison in this paper, this routing reaches the selected structured outputs with 19.0–$38.5\times$ **fewer updates** and **recognizable structure** 77–$100\times$ **earlier** than LoRA NJA.

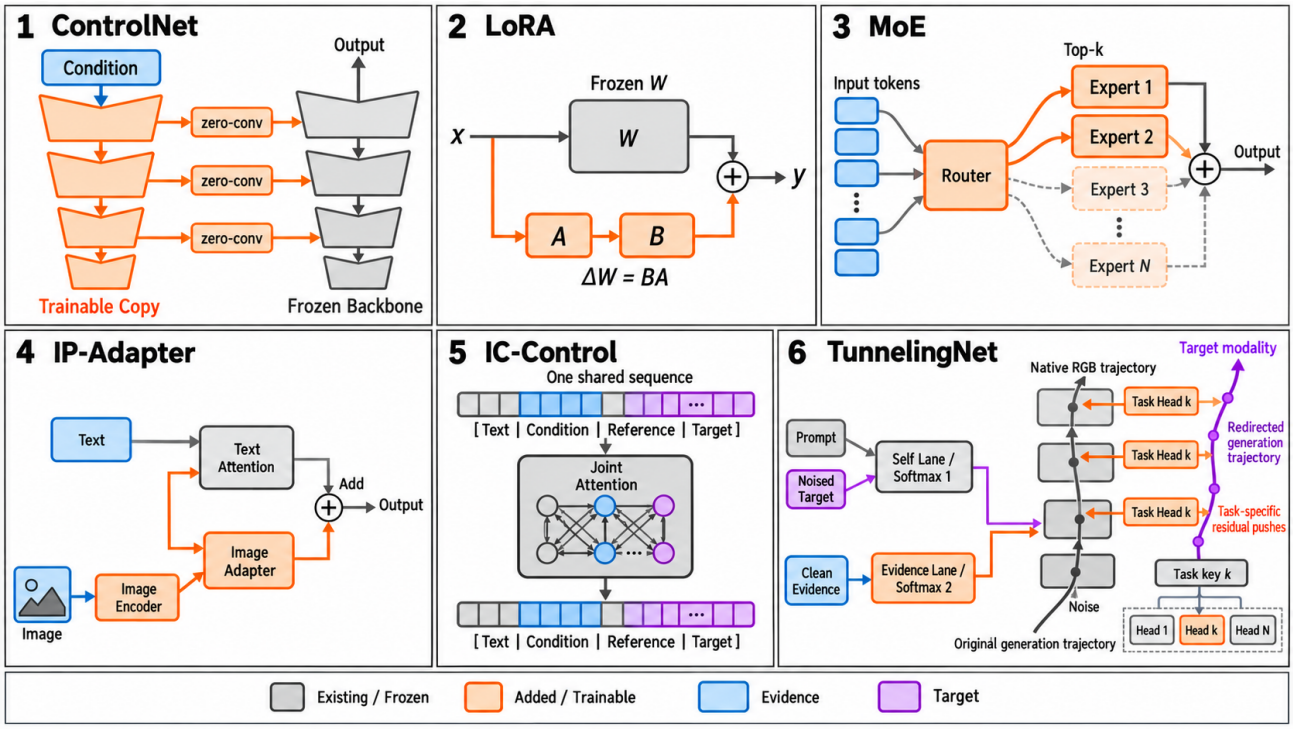


Figure 69. **A tutorial comparison of six paradigms for extending visual capabilities.** ControlNet adds a copied control body; LoRA adds low-rank weight updates; MoE selects expert computation; IP-Adapter adds an image conditioning branch; IC-Control-style methods organize conditions, references, and targets in a shared context; TunnelingNet applies task-specific residual pushes at multiple depths to redirect a frozen backbone's native RGB trajectory toward a selected target modality.

13.5. A Detailed Comparison of the Six Paradigms

The compact drawing in Figure 69 becomes more informative when each paradigm is examined along the same system dimensions. Table 16 reconstructs that comparison. It is not a ranking in which one method must replace all others: each established paradigm solves an important problem. The comparison identifies the additional organization required when the goal changes from adding one capability to building a large, directed, maintainable capability system.

Table 16. **Full tutorial comparison of six paradigms for extending visual capabilities.** The rows distinguish an architectural primitive from a complete task-system organization.

Dimension	ControlNet	LoRA	MoE	IP-Adapter	IC-Control style	TunnelingNet / WSM
Method level	Large auxiliary control architecture	Parameter-efficient weight update	Conditional-computation and capacity architecture	External image-conditioning adapter	Joint visual-context organization	Condition–target task architecture spanning representation, routing, trajectory deflection, training, and deployment
Intuition	Copy a large control body beside the backbone	Attach a small numerical patch to selected weights	Add experts and route tokens among them	Attach an image-prompt branch	Place text, conditions, references, and target in one context	Open a small target tunnel inside a shared backbone and steer its generation path repeatedly through depth
Main modification site	Copied multi-layer network with injections into the backbone	Weight-side branches of selected linear, attention, or feed-forward layers	Expert layers and routers	Image encoder, projector, and adapter attention	Input sequence, joint attention, and associated fine-tuning layers	Target-side residual paths plus a dedicated condition-to-target evidence route
Added structure	Near-backbone-scale control body and injection links	Low-rank matrices A and B	Expert networks and router	Image encoder, projector, and adapter attention	Context organization, usually with fine-tuning or LoRA	Task tunnel, task key, condition–target roles, interface type, and evidence routing
Backbone treatment	Usually frozen	Original weights usually frozen	Implementation dependent; experts and routing are coordinated	Usually frozen	Shared model or LoRA is commonly fine-tuned	Generative backbone, text encoder, and VAE are frozen
How conditions enter	A copied control body extracts features and injects them layer by layer	Uses condition interfaces already present in the host architecture	Router assigns tokens to experts	Independent image-conditioning branch enters attention	Condition and target tokens interact in a common context	Clean evidence acts on the noised target through a declared evidence route
Meaning of a visual task	Usually a new spatial condition for an RGB target	Defined by the training data; LoRA itself defines weight change, not an input–output task	Defined jointly by the shared objective and expert roles	Usually image prompt plus text to RGB	A contextual relation learned among jointly represented tokens	Every input–target mapping has its own generation trajectory, TunnelingNet, and task key
Information route	External control route through copied layers and injections	Network topology and token routes normally remain those of the host model	Router selects computation modules	Independent image-prompt route	Shared context provides joint interaction	Target self-generation is separated from condition evidence; the interface selects MTA or native joint attention
Attention resources	Control features are extracted before backbone injection	Under native joint attention, heterogeneous tokens share one Softmax budget	Tokens compete for or receive routing capacity	Text and image conditions commonly use separate attention before fusion	All tokens interact in the shared attention field	MTA separately normalizes target self-generation and condition evidence; NJA remains available when full interaction is useful
How output changes	Strong layer-wise control of RGB generation	Low-rank change to model function without explicit task routing	Selected experts provide different computation or knowledge	Image prompt augments text conditioning	Joint context learns relationships among all elements	Small task-specific residual forces progressively redirect the shared RGB trajectory into a selected target space
Unit of expansion	Copied control body and injection links	Set of low-rank weight differences	Expert capacity and routing	Adapter or conditioning interface	Contextual task organization and its training	Directed task edge jointly defined by conditions, target, interface, tunnel, and task key
Typical range	Depth, edge, or pose to RGB	Depends on data and objective	Depends on expert definitions	Image prompt and text to RGB	Text and visual contexts to a visual target	Directed generation among appearance, geometry, pose, motion, material, physical, semantic, sparse, and audio information

Continued on the next page

Table 16 continued

Dimension	ControlNet	LoRA	MoE	IP-Adapter	IC-Control style	TunnelingNet / WSM
Adding a capability	Adds another substantial control body	Adds an independent patch, but native information routes remain	May require coordination with experts and router	Adds an independent conditioning adapter	Changes the shared context and may require re-coordination	Trains and installs one task tunnel without retraining the backbone or existing routes
Capability composition	External workflow combines control modules	Switch, merge, or stack weight patches	Router selects or mixes experts	Combines image prompt with text and other controls	Combines evidence inside one context	One route's decoded output becomes clean evidence for the next declared route, creating executable visual chains
Scaling pressure	Large control structure per capability	Separate training and management; native evidence competition may remain	Expert count, routing, and coordinated-training cost	Primarily expands image conditioning, not arbitrary modality graphs	Sequence length, joint-interaction cost, and cross-modality interference	Small independently trainable tunnels; shared prior, task graph, and chains can expand while one request activates only selected routes
Principal strength	Mature spatial control	Efficient, general, distributable adaptation	Scalable conditional capacity	Strong image-prompt conditioning	Flexible contextual interaction	Unified task grammar, world-information representation, target routing, lifecycle management, and composable directed generation

13.6. What the Comparison Means

Put simply, natural-language processing now has large language models that handle **approximately 900 language tasks** within a common paradigm. Vision has **no comparable model**; existing visual methods generally unify **no more than approximately 30 tasks**. The main reason is that visual tasks are harder to train, use incompatible data formats, and optimize inconsistent objectives.

Training a model from scratch with dozens of input types and dozens of output types is impractical. LLMs succeeded in part because even code is structured language: linguistic data are inherently more consistent and easier to train on than visual data.

Computer vision has therefore largely progressed by adding new capabilities to existing image and video generation models, and the five paradigms above are among the most common and effective ways to extend such models.

We take a different approach: reaching hundreds or thousands of new targets by deflecting the generation trajectory of a shared backbone. Because visual information can be represented in image-like form, we call the resulting representation World Snapshot Graphics, or a World Frame, and the model a World Snapshot Model. Because the method reuses an existing generative backbone while opening routes to different targets, the analogy of creating new tunnels motivates the name TunnelingNet.

13.7. What Constitutes the Unified Capability

No single familiar ingredient is presented as the contribution in isolation. The essential construction is their organization into a scalable visual-task system:

- **A clear mapping** from conditional world information to target world information;
- **Non-RGB world states treated as actual generation targets**;
- **Independent addition, replacement, repair, and upgrade of task** capabilities;
- **Composable generation chains** among multiple visual tasks;
- separation between capability-library size and single-route active computation;
- **Target-side tunnels paired with condition-evidence routing**;
- **Separate processing of target generation and condition evidence** by MTA when the interface requires it; and
- **A large, flexibly composable mapping structure over visual tasks.**

This organization supports uncommon and bidirectional conversions whenever the modalities are sufficiently related, as well as many-to-one, one-to-many, and many-to-many contracts. RGB can also serve as additional evidence when a structural modality alone cannot determine texture, boundaries, material, or scene details. The resulting interface covers dense continuous fields, categorical maps, sparse structures, body representations, and audio tokens without requiring all targets to share one decoder.

Table 17. **Illustrative per-capability parameter scale relative to one 12.51M-parameter TunnelingNet.** The comparison uses representative configurations of each paradigm; implementations vary.

Extension paradigm	Relative scale
TunnelingNet	$1.0\times$
IP-Adapter	$\sim 1.8\times$
IC-Control style	$\sim 4\times$
LoRA	typically 1–$8\times$
ControlNet	$\sim 29\times$
MoE	~ 300–$10{,}000\times$
Full fine-tuning	$\sim 312.5\times$

The scale numbers in the main paper can now be read in the correct way. WSM v1 evaluates 1,148 routes on one shared visual backbone; it does not execute all routes simultaneously. Each dense route adds approximately 12.51M trainable parameters, **about 0.32% of the 3.9B visual backbone**. Its complete resident visual parameter bank is **approximately 18.54B parameters**, including all dense task tunnels and deployed categorical refiners. An ordinary dense request, however, activates **approximately 3.913B visual parameters**: the backbone and one selected tunnel. Copying a 3.9B backbone for every evaluated task would require **approximately 4.48T visual parameters**. Sharing the backbone therefore reduces this task-library parameterization by **about $241\times$** while retaining a dedicated route for every task. Capability-library size and per-request active computation are separate quantities.

13.8. How Parameter and Training Costs Scale with More Capabilities

The qualitative differences above become decisive when capabilities are added at scale. A dense TunnelingNet contains 12.51M trainable parameters. Under the representative configurations used in our engineering comparison, the per-capability parameter or weight-memory increment of established extension paradigms is approximately the multiple shown in Table 17. These ratios are configuration-dependent rather than universal constants, but they expose the different scaling coefficients of the six paradigms.

A 10,000-capability library makes the scaling difference concrete. Linearizing the representative increments in Table 17, TunnelingNet adds **approximately 125.1B resident parameters**, producing **a roughly 129B-parameter complete model** while an ordinary request still activates only the backbone and one selected TunnelingNet. The corresponding increments are about 225B for IP-Adapter,

500B for IC-Control, 125B–1.00T for LoRA, 3.63T for ControlNet, 37.5T–1.25P for MoE, and 39.1T for separate full fine-tuning. At this scale, most alternatives become difficult to train, store, transfer, and repeatedly upgrade. The conclusion is supported by measured experiments, not parameter arithmetic alone. Under matched trainable parameters, TunnelingNet reduces pooled training time by 20.6×; across five representative external tasks, WSM records **23 first-place and three second-place results**. The scaling advantage is therefore achieved without trading away task quality. At the ultra-large scale of hundreds to 10,000 capabilities considered here, WSM is the only evaluated approach that simultaneously maintains a manageable resident library, near-backbone active computation per request, rapid capability acquisition, and competitive or leading task quality. These properties make it the most practical currently demonstrated route toward ultra-large, LLM-scale visual capability systems containing hundreds of billions, and eventually trillions, of resident parameters.

Parameter count alone does not explain the acquisition advantage. Our parameter-matched LoRA NJA and TunnelingNet comparison holds the trainable parameter count fixed. Across Tasks 003, 275, and 313, the measured training time falls from 41.58 GPU-hours (1.73 GPU-days) to 2.02 GPU-hours (0.084 GPU-days), a pooled 20.6× acceleration. Applying the same measured per-task averages to 1,148 capabilities gives an illustrative capability-library acquisition cost of approximately 663 versus 32.2 GPU-days, a reduction of about 631 GPU-days. This extrapolation describes repeated independent capability acquisition; it is not substituted for the measured three-task result.

The scale becomes larger under industrial training schedules. SenseNova-Vision reports 50,000 training steps on at least 16 80GB GPUs [59], whereas the corresponding WSM task uses 1,740 single-GPU updates. Counting GPU-updates gives a conservative workload proxy of at least

$$\frac{50{,}000 \times 16}{1{,}740} \approx 460 \times .$$

If every one of 1,148 capabilities were acquired at that workload, the illustrative total would be **approximately 242,000 GPU-days**, compared with about **526 GPU-days** at the observed WSM task time, saving **approximately 241,000 GPU-days**. This is a transparent scale scenario based on published steps and GPU counts, not a measured SenseNova wall-clock or financial result.

Let c be GPU rental cost in dollars per GPU-hour and R the number of full capability-library acquisition or upgrade cycles. The corresponding GPU-rental-equivalent saving in this industrial scenario is

$$S(R, c) \approx 241{,}000 \times 24 \times c \times R.$$

At $c = \$2$–$\$4$, one complete cycle is approximately $11.6–$23.2 **million**; ten cycles are $116–$232 **million**; and one hundred hyperscale cycles are $1.16–$2.32 **billion**. The arithmetic explains why the same method can save only hundreds of thousands of dollars for a small deployment yet reach tens of millions per full training cycle, and hundreds of millions or billions under repeated hyperscale development. These estimates exclude data, annotation, engineering labor, storage, networking, and inference.

The same acceleration also changes an individual research cycle. Using the measured averages above, 50–100 training runs, including seeds, ablations, and hyperparameter trials, decrease from **approximately 29–58 to 1.4–2.8** single-GPU days. This saves **about 27–55 GPU-days**, or roughly $1,300–$5,300 at $2–$4 per GPU-hour, while allowing results to be inspected and experiments redirected within the same day.

Deployment exhibits a separate saving. Replicating a 3.9B backbone for 1,148 tasks requires approximately **4.48T parameters, or 8.96TB** of FP16/BF16 weights. The complete 18.54B-parameter WSM occupies **approximately 37.1GB of weights**, and an ordinary dense request activates **approximately 3.913B parameters, or 7.83GB** of weights. These are weight-only footprints; runtime memory additionally contains activations and framework state.

13.9. Capabilities That Emerge at the System Level

The table also reveals why the contribution cannot be reduced to another adapter or a collection of LoRAs. A weight-update primitive answers how a parameter is changed. A capability system must additionally specify what information is the condition, what world state is generated, how evidence reaches that target, how the route is selected, how routes compose, and how one capability changes over its lifetime. Once these relations are explicit, several system-level abilities become available together:

- **Uncommon, bidirectional, and multi-input/output contracts.** Related modalities can be connected in either direction, and contracts can be one-to-many, many-to-one, or many-to-many instead of being restricted to an external condition that controls RGB.
- **Executable visual chains.** A decoded result becomes clean evidence for the next route. This is structurally different from merging weights or selecting experts because the intermediate world state and the next target remain explicit and inspectable.
- **Independent capability lifecycle.** A route can be trained in parallel, installed, tested, repaired, replaced, upgraded, or rolled back without changing the shared backbone or previously registered routes.
- **Target-specific information flow.** Only the target-side tunnel is optimized. MTA can give target self-generation and condition evidence separate normalization, while native joint attention remains selectable for interfaces that benefit from full interaction.

- **Heterogeneous native outputs.** Dense continuous fields, categorical maps, keypoints, boxes, skeletons, body parameters, and audio tokens coexist in one task system without pretending that every output has the same decoder or constraint.
- **RGB-assisted recovery.** RGB can accompany an incomplete structural condition when texture, boundaries, material, identity, or scene detail cannot be recovered from that condition alone, creating a distinct family of RGB-assisted inverse-rendering contracts.

This organization also creates visual tasks instead of merely collecting known ones. Among 1,072 audited cross-variable contracts, **at least 655** have no clear historical input–output counterpart; including historically underspecified cases gives an **upper estimate of 1,056**. The result is a system that both covers the established task inventory and exposes previously unnamed relationships among world properties.

13.10. A Practical Mental Model of WSG, WSM, and TunnelingNet

The three names describe different levels of the system.

- **WSG is the shared representation.** RGB occupies channels 0–2, while additional channel IDs store other world information. Readers and adapters expose dense maps, categories, keypoints, boxes, skeletons, body parameters, and audio-related targets through their native forms.
- **WSM is the shared generative system.** It contains a frozen visual prior and a library of directed capabilities. It can be understood as a computer with many specialized applications installed: all applications share the same operating system, and opening one application activates only the selected application instead of executing the entire library.
- **TunnelingNet opens and executes a route.** A task key selects the required target tunnel and evidence topology. The tunnel changes the generation path at multiple depths; it does not require another copy of the shared backbone.

This mental model explains why installing more capabilities need not increase the active computation of an ordinary single-route request. A new task trains only its target-side tunnel, leaving the backbone and existing tunnels unchanged. Different tasks can be trained independently and in parallel. After training, a capability can be installed, tested, repaired, replaced, upgraded, or rolled back without retraining the other routes.

Adding one capability. Training a new capability has five steps: (1) prepare paired conditions and targets for the desired conversion; (2) register the target's WSG channel identity and native reader; (3) select the interface, evidence roles, and task key; (4) freeze the shared backbone, text encoder, VAE, and existing routes while training the target tunnel; and (5) validate, version, and install the resulting capability. A later correction repeats only the affected route's training and validation. This workflow is why the system can acquire capabilities on different GPUs in parallel and why updating one route does not require a new joint training cycle for the complete library.

The same organization also explains the role of visual chains. Switching, merging, or stacking weight patches combines parameters, but it does not by itself define an executable sequence among visual world states. In WSM, the output of one directed task can become clean evidence for the next. A sequence such as

$$\mathrm{RGB} \rightarrow \mathrm{Depth} \rightarrow \mathrm{Normal} \rightarrow \mathrm{Edge}$$

is therefore a chain of explicit visual contracts. Each step has a declared target and can be inspected, replaced, or reused independently.

13.11. From a Route to a Visual Capability Operating System

WSM can be viewed as a computer with many specialized applications installed. The shared backbone plays the role of a common operating system; each registered route supplies one application-level capability; and a task key opens only the capability requested by the user. Installing another capability expands the library but does not cause every installed route to execute on every request.

ControlWorld turns this model into an end-user workflow. A user selects the condition and target in the task graph, supplies image, text, reference, or world-information evidence, executes the corresponding tunnel, views the result through the target's reader, and may pass that result into the next route. This supports workflows such as

- RGB → depth → normal → edge;
- RGB → roughness → base color;
- non-RGB conversions such as normal → edge;
- edge, depth, normal, or pose → RGB generation;
- first-frame appearance plus current structure → a new structural or appearance state; and
- structural evidence plus RGB → geometry, material, semantics, pose, or other world properties.

The engineering difference is direct. A conventional pipeline may maintain a separate depth model, normal model, edge model, pose model, and RGB generator, plus conversion code between their incompatible formats. WSG gives intermediate world states stable identities, the shared backbone supplies a common prior, and task tunnels provide replaceable conversions. The output of one step can be reused immediately instead of being treated as an accidental by-product of an opaque end-to-end model.

13.12. Explicit Workflows and an Embedded Task Graph

The task graph is a semantic organization, not a requirement that the system must appear as a visible workflow or as thousands of unrelated model files. In ControlWorld, the graph is exposed so that a user can select, inspect, and chain routes. In a deployed WSM, the same task keys, condition–target roles, tunnel parameters, and readers can reside inside one model package and be selected by the request. Physical sharding is a storage decision; it does not define the boundary of the method or the model.

This distinction also separates *capability acquisition* from *capability execution*. Routes have different targets and can therefore be learned independently, even on different GPUs, without placing all modalities, losses, and targets in one joint training batch. At inference, only the route or route chain named by the task key is activated. The library can consequently grow through distributed acquisition while retaining one shared visual prior and predictable per-request computation.

An additional research direction follows from this organization. A self-contained WSM may use one of its own estimated world states as evidence for a later route, or use an audited chain to construct training signals for a new route. This is not required by the present evaluation; it is a path toward recursive visual systems whose intermediate reasoning remains represented as inspectable world information instead of hidden activations alone.

This organization points beyond a single benchmark. The same capability library can support richer perception for robotics and digital twins; geometry, motion, material, lighting, and rendering chains for film and games; pose, shape, and segmentation routes for digital humans, health, and sports; physical and material property estimation for manufacturing and logistics; editable product assets for e-commerce; audiovisual localization and segmentation; and institution-specific routes for medical, agricultural, satellite, or scientific imagery. These are not separate claims that one frozen release already solves every domain. They illustrate how a common task grammar and independently trainable routes turn domain expansion into capability installation instead of complete-system reconstruction. The measurable quantities are correspondingly concrete: trainable parameters and updates per new route, time to recognizable structure, resident library size, active parameters per request, number of independently replaceable capabilities, and the length and success rate of executable chains.

13.13. World Snapshot Model Application Landscape

The World Snapshot Model framework opens a broad application landscape across foundational vision, capability expansion, generation, world-state computing, scientific and industrial systems, and visual-AI infrastructure. The following forward-looking map organizes 882 prospective application scenarios into 19 families. It serves as a research and development roadmap for the task graph, capability library, and world-information interfaces introduced above.

1. Foundational Computer Vision, World Intelligence, and Reasoning (38)

General-purpose computer vision; vision foundation models; multimodal vision foundation models; vision agents; multimodal vision agents; visual reasoning; visual chain-of-thought reasoning; compositional visual reasoning; large-scale visual reasoning; multimodal visual reasoning; cross-modal visual reasoning; visual physical reasoning; visual spatial reasoning; visual temporal reasoning; visual causal reasoning; visual counterfactual reasoning; visual commonsense reasoning; interactive visual-world reasoning; unified visual tasks; any-input, any-output vision systems; multi-input, single-output vision systems; multi-input, multi-output vision systems; open-world visual learning; automated visual task discovery; visual world-interaction models; interactive visual world models; persistent visual world models; dynamic visual world models; predictive visual world models; physical visual world models; in-cabin visual perception; visual search; video summarization; visual disaster damage assessment; cross-view verification; cross-temporal verification; cross-modal visual verification; *pixel-level style transfer.

2. Model Capability Expansion, Composition, and Runtime Management (37)

Multi-task visual learning; continual visual learning; incremental visual learning; modular visual learning; compositional visual learning; visual capability modules; hot-loading of model capabilities; hot updates of model capabilities; hot-swapping of model capabilities; incremental deployment of model capabilities; automated composition of visual capabilities; visual task routing; vision model routing; visual tool routing; visual expert routing; visual capability orchestration; world-capability orchestration; visual workflow optimization; visual workflow execution; visual workflow verification; vision model fusion; visual capability fusion; vision model distillation; visual capability distillation; vision model merging; vision model compression; visual capability versioning; visual capability rollback; visual capability diagnostics; visual capability verification; visual capability compatibility testing; automated expansion of visual capabilities; compositional expansion of visual capabilities; linear expansion of visual capabilities; nonlinear expansion of visual capabilities; homogeneous expansion of visual capabilities; heterogeneous expansion of visual capabilities.

3. Generation, Editing, Restoration, and Cross-Domain Translation (97)

Generative visual perception; generative visual tasks; multi-condition visual generation; multi-target visual generation; automated visual task generation; visual workflow generation; visual capability repair; general-purpose image generation; general-purpose video generation; long-form video generation; real-time video generation; interactive video generation; multiview image generation; multiview video generation; 3D content generation; 4D content generation; spatial video generation; controllable image generation; controllable video generation; reference-guided image generation; scene-consistent generation; object-consistent generation; physically consistent generation; temporally consistent generation; illumination-consistent generation; motion-consistent generation; audio-visual consistent generation; image editing; video editing; 3D editing; 4D editing; scene editing; object editing; illumination editing; motion editing; camera editing; facial-expression editing; apparel editing; background editing; localized region editing; semantic editing; interactive editing; generative restoration; image restoration; video restoration; 3D restoration; old-photo restoration; digital restoration of cultural heritage; image denoising; video denoising; image deblurring; video deblurring; image super-resolution; video super-resolution; image completion; video completion; scene completion; object completion; missing-modality completion; image colorization; video colorization; image style transfer; video style transfer; domain translation; cross-domain generation; image-to-image translation; video-to-video translation; audio-to-image generation; audio-to-video generation; image-to-audio generation; video-to-audio generation; text-to-image generation; text-to-video generation; text-to-3D generation; image-to-3D generation; video-to-3D generation; 3D-to-image generation; 3D-to-video generation; World Snapshot generation; World Snapshot video generation; World Snapshot field generation; world-state generation; future-state generation; world-state restoration; world-state completion; generative world models; apparel generation; media content generation; general-purpose generative interfaces; *unified 1D-to-4D generation; *text-to-world generation; *single-pass generation of tens to hundreds of world attributes; *World Snapshot fine-tuning of pretrained generative models; *auxiliary-attribute expansion for generated content; *pixel-level motion transfer; *multilayer human generation; *joint generation of human surface and internal structures.

4. World Snapshots, World States, and World Data (38)

World Snapshot Models; world-state prediction; world-

state updating; world-state versioning; world-state branching; world-state rollback; world-state querying; world-state fusion; world-state compression; world-state transmission; world-state databases; world-state search engines; world-state knowledge bases; world-data platforms; world-state digital assets; world-attribute interfaces; World Snapshot APIs; world-state APIs; *World Snapshot computer graphics; *World Snapshot line representation; *World Snapshot illumination representation; *World Snapshot shape representation; *latent-space entity representation; *World Snapshot graphics data acquisition; *World Snapshot video data acquisition; *World Snapshot field data acquisition; *world-information acquisition from machinery; *World Snapshot data mining; *World Snapshot exploration by agents; *PNG-to-WSG conversion; *WSG information extraction; *WSG storage; *WSG compression; *World Snapshot database systems; *secure World Snapshot computing; *efficient World Snapshot computing architectures; *cross-platform World Snapshot video players; *World Snapshot software engineering.

5. Recognition, Localization, Detection, Segmentation, and Tracking (67)

Generative visual understanding; visual completion of occluded regions; visual world-state retrieval; image classification; video classification; object classification; scene classification; visual action classification; visual behavior classification; visual event classification; audio-visual cross-modal classification; open-vocabulary image classification; fine-grained image classification; zero-shot image classification; few-shot image classification; image localization; object localization; keypoint localization; text-guided visual grounding; vision-language grounding; audio-visual sound-source localization; 3D visual localization; visual spatiotemporal localization; visual indoor localization; visual outdoor localization; object detection; open-vocabulary object detection; zero-shot object detection; few-shot object detection; 3D object detection; video object detection; small-object detection; dense object detection; occluded-object detection; visual anomaly detection; video event detection; video action detection; audio-visual sound-event detection; audio-guided visual instance detection; semantic segmentation; instance segmentation; panoptic segmentation; part segmentation; human segmentation; object segmentation; video segmentation; 3D segmentation; 4D segmentation; interactive segmentation; open-vocabulary segmentation; audio-guided visual instance segmentation; motion segmentation; scene segmentation; visual forest monitoring; visual vessel detection; video retrieval; video understanding; video question answering; visual question-answering assistance; image captioning; video captioning; visual object recognition; visual object tracking; visual terrain understanding; visual causal analysis; *occlusion reasoning for overlapping objects; *attribute inference for occluded regions.

6. Visual Geometry, Motion, Material, and Physical-Property Estimation (96)

Visual Model Deployment at the Edge; Geometry-Consistent Generation; Material-Consistent Generation; Material Editing; Pose Editing; Visual Physical-Property Editing; Visual Scene-State Estimation; Visual Material Classification; Visual Material Segmentation; Visual Affordance Segmentation; Visual Contact-Region Segmentation; Visual Risk-Region Segmentation; Depth Estimation; Monocular Depth Estimation; Multi-View Depth Estimation; Video Depth Estimation; Metric Depth Estimation; Relative Depth Estimation; Disparity Estimation; Surface Normal Estimation; Curvature Estimation; Edge Estimation; Contour Estimation; Optical Flow Estimation; Scene Flow Estimation; Visual Motion Estimation; Camera Pose Estimation; Object Pose Estimation; Human Pose Estimation; Hand Pose Estimation; Facial Pose Estimation; Skeleton Estimation; Visual Human-Parameter Estimation; Visual Object-Structure Estimation; Visual Scene-Structure Estimation; Visual Geometry Estimation; Visual Scale Estimation; Visual Distance Estimation; Visual Velocity Estimation; Visual Acceleration Estimation; Visual Trajectory Estimation; Visual Deformation Estimation; Visual Articulation-State Estimation; Visual Occlusion Estimation; Visual Visibility Estimation; Visual Illumination Estimation; Visual Reflectance Estimation; Visual Material Estimation; Visual Texture Estimation; Visual Shadow Estimation; Visual Transparency Estimation; Visual Roughness Estimation; Visual Metallicity Estimation; Visual Refractive-Index Estimation; Visual Density Estimation; Visual Mass Estimation; Visual Volume Estimation; Visual Weight Estimation; Visual Support-Area Estimation; Visual Friction-Coefficient Estimation; Visual Friction-Force Estimation; Visual Elasticity Estimation; Visual Stiffness Estimation; Visual Hardness Estimation; Thermal-Image Temperature Estimation; Visual Pressure Estimation; Visual Stress Estimation; Visual Strain Estimation; Visual Flow-Rate Estimation; Visual Contact-State Estimation; Visual Collision-Risk Estimation; Visual Affordance Estimation; Visual Stability Estimation; Visual Load-Bearing-Capacity Estimation; Visual Physical-Property Estimation; Visual Material-Property Estimation; Visual Uncertainty Estimation; Visual Confidence Estimation; Visual Risk Estimation; Visual Future-State Estimation; Visual Pedestrian-Trajectory Prediction; Visual Collision Prediction; Visual Geometry-Consistency Verification; *Wave-Function Visual Generative Models; *Native Visual Computation of World Physical Parameters; *Native Visual Modeling of Relative Distance; *Native Visual Modeling of Absolute Distance; *Visual Relative-Volume Computation; *Visual Relative-Mass Prediction; *Pixel-wise Mass-Distribution Modeling; *Multi-Wave-Function

Distributional Representations of Visual Worlds; *Attaching Visual Physical Properties to Generated Content; *Pose- and Motion-Conditioned Visual World Classification; *Depth- and Audio-Conditioned Visual World Classification; *Multi-Depth-Layer Visual Occlusion Representation; *Probabilistic Visual World-Line Prediction.

7. Reconstruction, Digital Twins, Simulation, and Scientific Computing (55)

Past Visual-State Reconstruction; Visual Digital Twins; Generative Visual Digital Twins; Real-Time Visual Digital Twins; Urban Visual Digital Twins; Human Visual Digital Twins; Image Reconstruction; Video Reconstruction; 3D Reconstruction; 4D Reconstruction; Multi-View Reconstruction; Dynamic Scene Reconstruction; Human Reconstruction; Hand Reconstruction; Face Reconstruction; Object Reconstruction; Indoor Scene Reconstruction; Outdoor Scene Reconstruction; Urban Visual Reconstruction; Roadway Visual Reconstruction; Terrain Visual Reconstruction; Subsurface-Structure Visual Reconstruction; Microscopy Image Reconstruction; Magnetic Resonance Image Reconstruction; Computed Tomography Image Reconstruction; Ultrasound Image Reconstruction; Neural Radiance Field Reconstruction; Gaussian Splatting Reconstruction; Point-Cloud Reconstruction; Mesh Reconstruction; Voxel Reconstruction; Occupancy-Field Reconstruction; Material Reconstruction; Illumination Reconstruction; Acoustic-Field Visualization and Reconstruction; Physical-Field Visualization and Reconstruction; Generative Visual Reconstruction; Generative RGB-D Reconstruction; Generative Multi-Attribute Reconstruction; Visual Material Representation; Visual Physics Simulation; Visual Fluid Simulation; Visual Structural Simulation; Visual Thermodynamic Simulation; Visual Electromagnetic-Field Simulation; Visual Acoustic-Field Simulation; Microscopy Vision; Audiovisual Scene Understanding; Visual Sim-to-Real Transfer; Visual Real-to-Sim Reconstruction; Generative Reconstruction APIs; *Visual Modeling of Dynamic Physical Fields; *Billion-Scale Natural Simulation-Scene Generation; *Recording World Information with Gaussian Representations; *Fusion of Gaussian Fields and World Snapshots.

8. Robotics, Embodied AI, and Autonomous Systems (48)

Vision-Based Embodied Intelligence; Robotic Visual Digital Twins; Visual Robot Localization; Robotic Visual World Models; Embodied Visual World Models; Vision-Language-Action Models; Robotic Visual Perception; Vision-Based Robot Planning; Vision-Based Robot Control; Vision-Guided Robotic Grasping; Vision-Guided Robotic Manipulation; Vision-Based Robot Navigation; Vision-Based Robot Obstacle Avoidance; Visual SLAM for Robotics; Robotic Visual Scene Understanding; Vision-Based Robot Task Planning; Visual Robot Skill Learning; Visual Robot Skill Composition; Visual Robot Imitation Learning; Visual Robot Reinforcement Learning; Vision-Based Autonomous Robot Exploration; Vision-Based Robot Fault Diagnosis; Robotic Visual-State Prediction; Visual Safety Control for Robots; Visual Multisensor Fusion for Robotics; Visual Human-Robot Collaboration; Home-Service Robot Vision; Industrial Robot Vision; Logistics Robot Vision; Warehouse Robot Vision; Aerial Robot Vision; Space Robot Vision; Biomimetic Robot Vision; General-Purpose Robot Vision; Vision-Based Drone Navigation; Vision-Based Drone Obstacle Avoidance; Visual Navigation-Map Generation; Robotic Vision for Space Stations; Vision-Based Autonomous Satellite Navigation; Vision-Guided Intraoperative Navigation; Marine Vessel Visual Navigation; Visual Navigation for Blind and Visually Impaired People; Military Robot Vision; Visual Search-and-Rescue Robotics; Home-Robot Visual Control; Vision-Based Multi-Robot Collaboration; Robotic Vision Operating-System Extensions; Embodied Visual Intelligence APIs.

9. Autonomous Driving, Transportation, Drones, and Aerospace (39)

Vehicle Visual Digital Twins; Visual Vehicle Localization; Visual World Models for Autonomous Driving; Vision-Based Autonomous Driving; Vision-Based Driver Assistance; Vehicle Visual Perception; Vision-Based Vehicle Planning; Vision-Based Vehicle Control; Lane Detection; Road Segmentation; Drivable-Area Detection; Traffic-Participant Detection; Visual Traffic-Scene Prediction; Visual Vehicle-Trajectory Prediction; Visual Driving-Risk Assessment; Visual Driver-Behavior Understanding; Vision-Based Driving Decision-Making; Vision-Based Parking Assistance; Vision-Based Automated Parking; Visual Driver Monitoring; Cooperative Vehicle-Infrastructure Vision; Traffic Visual Digital Twins; Visual Traffic Simulation; Visual Data Generation for Autonomous Driving; Autonomous-Driving Scene Reconstruction; Visual Counterfactual Testing for Autonomous Driving; Vision-Based Drone Inspection; Vision-Based Drone Mapping; Vision-Based Drone Formation Control; Visual Object Tracking by Drones; Drone-Based Visual Environment Perception; Aerial Image Analysis; Satellite Image Analysis; Astronomical Image Analysis; Telescope Image Analysis; Planetary-Surface Image Analysis; Visual Space-Domain Awareness; Visual Search-and-Rescue Drones; Visual Traffic-Accident Analysis.

10. Remote Sensing, Earth, Cities, Architecture, and Infrastructure (46)

Building Visual Digital Twins; Earth Visual Digital Twins; Environmental Visual Digital Twins; Visual Geolocation; Remote-Sensing Image Segmentation; Visual Building Reconstruction; Remote-Sensing Image In-

terpretation; Remote-Sensing Land-Cover Classification; Remote-Sensing Land-Use Analysis; Visual Disaster Monitoring; Visual Flood Monitoring; Visual Fire Monitoring; Visual Wildfire Prediction; Image-Based Earthquake-Damage Assessment; Visual Landslide Monitoring; Visual Glacier Monitoring; Visual Coastline Monitoring; Visual Water-Body Monitoring; Remote-Sensing Environmental Change Detection; Visual Weather Analysis; Visual Geospatial Reasoning; Map Generation; Visual Map Updating; 3D Mapping; High-Definition Map Generation; Indoor Visual Mapping; Visual Monitoring of Urban Infrastructure; Visual Road Inspection; Visual Bridge Inspection; Visual Tunnel Inspection; Visual Railway Inspection; Visual Airport Inspection; Visual Port Inspection; Visual Power-Grid Inspection; Visual Pipeline Inspection; Visual Water-Infrastructure Inspection; Visual Telecommunications-Infrastructure Inspection; Visual Building Information Modeling; Visual Construction-Site Monitoring; Visual Building-Quality Inspection; Visual Building-Energy Analysis; Visual Spatial-Layout Generation; Visual Furniture-Layout Generation; Visual Real-Estate Presentation; Virtual Property Tours; Architectural Visualization.

11. Industrial, Manufacturing, Logistics, Retail, and Commercial Applications (37)

Industrial visual digital twins; Automated visual defect detection; Industrial defect segmentation; Machine vision; Automated visual inspection; Surface defect detection; Visual dimensional inspection; Visual assembly inspection; Visual weld inspection; Visual part recognition; Visual part pose estimation; Visual production-process monitoring; Vision-based equipment condition monitoring; Vision-based predictive maintenance; Visual fault localization; Visual root-cause analysis; Production-line visual digital twins; Supply-chain visualization; Vision-based warehouse automation; Vision-based logistics sorting; Visual package inspection; Vision-based inventory recognition; Visual shelf monitoring; Visual product tracking; Retail computer vision; Visual product recognition; Visual product search; Shelf-image analysis; Virtual try-on; Product image generation; Advertising image generation; E-commerce visual content generation; Brand visual asset generation; Marketing visual content generation; Industrial visual design; Industrial visual safety monitoring; Image-based insurance damage assessment.

12. Medicine, Life Sciences, Agriculture, Forestry, and Marine Applications (48)

Medical visual digital twins; Medical image segmentation; Underwater visual reconstruction; Medical image reconstruction; Medical robot vision; Surgical robot vision; Rehabilitation robot vision; Agricultural robot vision; Underwater robot vision; Medical image analysis; Image-assisted clinical diagnosis; Lesion detection in medical images; Tumor segmentation in medical images; Organ segmentation in medical images; Anatomical reconstruction from medical images; Image-guided surgical planning; Image-guided radiotherapy planning; Disease-progression prediction from medical images; Longitudinal medical image analysis; Histopathology image analysis; Digital pathology; Cell image segmentation; Cell tracking in microscopy images; Tissue-structure analysis; Protein imaging analysis; Vision-based biomechanical analysis; Vision-based human motion analysis; Vision-based rehabilitation assessment; Vision-based vital-sign estimation; Healthcare visual digital twins; Agricultural computer vision; Vision-based crop monitoring; Visual pest and disease detection; Visual weed detection; Visual fruit counting; Visual fruit-ripeness estimation; Vision-based crop-yield prediction; Vision-based soil-condition estimation; Agricultural drone vision; Vision-based livestock monitoring; Vision-based aquaculture monitoring; Marine computer vision; Underwater scene understanding; Underwater object detection; Underwater 3D reconstruction; Vision-based marine ecosystem monitoring; Vision-based fisheries monitoring; *World Snapshot-based medical information recording.

13. Film, Media, Games, XR, Design, and Culture (64)

Immersive visual content generation; Identity-consistent character generation; Cross-shot visual character consistency; Visual character editing; Game visual world models; Film and television visual world models; AI-generated films; AI-generated television series; AI-generated animation; AI-generated short-form video; AI-powered visual live streaming; Virtual production; Visual previsualization; Visual shot planning; Visual storyboard generation; Visual shot-sequence generation; Visual character generation; Digital visual actors; Visual virtual characters; Visual digital humans; Visual virtual presenters; Vision-based motion capture; Vision-based facial-expression capture; Visual facial reenactment; Visual lip synchronization; Audio-driven visual animation; Music visualization; Visual effects; Visual post-production; Automated video editing; Visual media asset management; Visual game generation; Game-world visual generation; Visual game-level generation; Visual game-character generation; Visual game-animation generation; Visual game-material generation; Visual game-physics generation; Visual reconstruction of game scenes; Vision-based game agents; Visual intelligence for non-player characters; Visual interactive storytelling; Persistent visual game worlds; Open visual game worlds; Live visual-content updates for games; Visual game-mod ecosystem; Visual cloud gaming; Virtual-reality visual generation; Augmented-reality visual generation; Mixed-reality visual generation; Visual spatial computing; Virtual visual laboratories; Visual content for virtual museums; Dig-

ital visual exhibitions; Visual digitization of cultural heritage; Visual archaeological reconstruction; Artwork image analysis; AI visual art creation; AI-assisted visual design; AI-assisted visual product design; AI-assisted visual graphic design; *World Snapshot film generation; *Two-hour feature-length film generation; *Infinite game-frame generation.

14. Education, Sports, Sign Language, and Accessibility (22)

Vision-based teaching; Visual educational content generation; Sports computer vision; Vision-based athlete pose analysis; Sports video analysis; Tactical analysis of sports video; Vision-based movement scoring; Vision-assisted officiating; Vision-based fitness coaching; Golf video analysis; Soccer video analysis; Basketball video analysis; Tennis video analysis; Swimming video analysis; Dance video analysis; Visual sign-language recognition; Sign-language video generation; Real-time visual sign-language translation; Assistive vision; Visual assistance for people with low vision; Accessible visual content generation; Video caption generation.

15. Public Safety, Identity Security, Defense, and Emergency Response (36)

Public-safety video analytics; Video surveillance; Abnormal-behavior detection in video; Person tracking in video; Video crowd analysis; Video crowd-density detection; Unattended-object detection in video; Video intrusion detection; Visual border surveillance; Airport visual security; Campus visual security; Visual construction-site safety monitoring; Vision-based fire early warning; Visual smoke detection; Visual hazardous-material detection; Visual identity verification; Face recognition; Gait recognition; Visual liveness detection; Forensic image and video analysis; Image authenticity detection; Video authenticity detection; Deepfake detection; Visual content provenance; Visual content copyright protection; Privacy-preserving computer vision; Defense computer vision; Visual battlefield situational awareness; Visual intelligence, surveillance, and reconnaissance; Cooperative unmanned vision systems; Vision-based equipment condition monitoring; Visual disaster-response analysis; Visual reconstruction of disaster-affected areas; Visual casualty localization; Visual accident reconstruction; Visual scene reconstruction.

16. Scene-Centric Vision, Edge Devices, and Multisensor Fusion (28)

Multisensor visual detection; Visual understanding of home environments; Visual digital twins of home environments; Vision-based elder care; Vision-based child care; Vision-based pet monitoring; Vision-enabled smart glasses; Wearable vision; Mobile vision; Edge vision; IoT vision; Visual sensor fusion; Radar-camera fusion; LiDAR-camera fusion; Visuotactile fusion; Audiovisual fusion; Visual-inertial fusion; Multispectral fusion; Hyperspectral fusion; Thermal-visible fusion; Event-camera fusion; Cross-sensor visual generation; Vision-based missing-modality completion; Visual sensor calibration; Visual sensor simulation; *Distance-attenuated audio in visual space; *Location-dependent sound generation in explorable visual worlds; *Visual learning of radar-derived range cues.

17. Data Production, Evaluation, Trustworthiness, and Collaborative Learning (28)

Synthetic data generation; Data augmentation; Automated annotation; Pseudo-label generation; Training-data generation; Evaluation-data generation; Benchmark generation; Long-tail data generation; Rare-event generation; Safety-testing data generation; Counterfactual data generation; Model evaluation; Model diagnostics; Model interpretability; Visual consistency verification; World-attribute consistency verification; Cycle-consistency verification; Physical-consistency verification; Semantic-consistency verification; Automated repair of generated outputs; Multi-model collaboration; Multi-agent collaboration; Visual-agent collaboration; World-state sharing; Distributed world models; Federated visual learning; Privacy-preserving multimodal learning; Cloud-edge collaborative vision.

18. Platforms, Operating Systems, Interfaces, APIs, and Capability Ecosystems (28)

Visual capability marketplace; Visual plug-in ecosystem; Vision capabilities as a service; Visual tasks as a service; Multimodal capabilities as a service; Visual model serving; Cloud platform for visual models; Private deployment of visual models; Multi-tenant visual model services; Visual capability licensing; Visual capability metering; Visual capability trading; Visual data platform; Multimodal data platform; Real-time vision services; Large-scale visual task platform; Physical AI platform; Spatial intelligence platform; World simulation platform; Universal world interface; Universal visual interface; World capability API; Visual reasoning API; Compositional vision API; Visual capability plug-in marketplace; World-model plug-in marketplace; Enterprise multimodal platform; Ultra-large-scale visual task services.

19. Visual AI Engines, World Engines, and Generative Rendering (30)

General-purpose visual AI engine; World Snapshot AI engine; Visual foundation-model engine; Multimodal visual AI engine; Visual generation engine; Visual reasoning engine; Compositional visual reasoning engine; Visual world-reasoning engine; Embodied visual intelligence engine; Robotic visual intelligence engine; Autonomous-driving visual intelligence engine; Physical visual intelligence engine; Spatial visual intelligence engine; Visual capability engine; World capability engine; Interactive-world

vision engine; World-simulation vision engine; World-state vision engine; Computer graphics engine; Generative visual rendering engine; Neural rendering engine; Real-time rendering engine; Physics-based rendering engine; Differentiable rendering engine; 3D scene rendering engine; 4D dynamic-scene rendering engine; Film and television rendering engine; Game rendering engine; XR rendering engine; Digital-twin rendering engine.

14. Memoir: Experimental Derivation of the Unified Law

This section tells how we arrived at the Unified Law. In 2024, we began with a simple idea: if an image could carry appearance together with other information about the world, image processing could become world-information processing. Over the following two years, this idea grew into a large experimental program: **60 experiment families**, at least **198 explicitly named methodological variants**, and **at least 2,293 training attempts documented in surviving records**. Among them, **1,546 different training runs** retain complete configurations, and at least **1,380 trained models** retain their weights. These counts are conservative because some older experiments were deleted, moved, or archived elsewhere. We first expanded image codecs, then built translators between latent spaces, then studied joint multimodal training, and finally moved the conversion into the generation process itself. Each stage answered one question and led naturally to the next.

The experiments covered different model families, training objectives, and foundation models. Their value is therefore larger than a comparison between a few network blocks. Together they showed where visual-task unification becomes simple, where it becomes difficult, and which recurring principles support the final Unified Law. These were not blind architecture sweeps. Nearly every variant began with a concrete hypothesis raised by the preceding result. The account below follows the development in order, grouping adjacent variants that tested the same hypothesis into at least 60 experiment families, and names at least 198 substantive methodological variants. Counting changes in network size, objective, schedule, path construction, and integration depth raises the total further; random seeds, repeated runs, and routine hyperparameter sweeps are excluded.

14.1. We First Tried to Put All World Information into One Codec

Our starting point was WSG. Standard image generation produces RGB. We imagined an image-like file whose first channels store RGB and whose additional channels store depth, normal, material, motion, pose, segmentation, and other world properties. If one model could encode and decode this complete file, many visual tasks could share one representation immediately.

Our first hypothesis was that a strong RGB VAE [69] might already contain most of the visual knowledge we needed, leaving only the additional world information to be inserted. We therefore enlarged its input and output layers to receive new channels. We pictured this as adding another layer outside an existing pyramid: the mature RGB codec would remain inside, while the new layer would carry world information into and out of its latent space. Directly training the enlarged codec quickly disturbed the original RGB space, so we froze the RGB path and trained only the new outer route. A LayerDiffuse-style auxiliary encoder [67] made this substantially more stable and showed that extra information could travel beside RGB. It also exposed the limit of that design. If the new route was kept weak, decoded properties became smooth averages; if it was made strong, it disturbed the RGB space it was meant to preserve. Newly initialized codecs, continuous latents, discrete VQ representations [70], and special handling for sparse signals tested whether another boundary could remove this tradeoff.

The first implementation expanded a pretrained TinyVAE progressively. Its hidden width grew from the original 64 channels in steps of 16, while its latent width grew from four channels to five, six, seven, and eight. Each stage froze the preceding RGB path and trained the newly added capacity on the full 256-channel WSG sample. This experiment tested a very direct hypothesis: if the RGB codec already described the visible world, a small amount of new capacity might be enough to attach the remaining properties. Dense fields such as depth began to reconstruct, while sparse pose channels did not. The full-channel tensor also made every update far slower than the original TinyVAE. Progressive widening preserved an existing codec, but it did not make all information types equally easy to insert.

We then studied where the additional field should enter the frozen codec. The first offset design encoded RGB together with the requested WSG field, predicted an offset Δz, and formed an adjusted latent $z' = z_{\mathrm{RGB}} + \Delta z$. This was the most direct way to make the new field travel with an existing RGB latent, but it also changed the very tensor that the pretrained decoder used to reconstruct RGB. A strong offset damaged the pretrained appearance distribution; a weak offset preserved RGB but carried too little information about the new field. This experiment made the central tension visible before we changed any decoder architecture.

We next held the adjusted latent fixed and changed where its new information entered the decoder. The first version injected the offset at one intermediate decoder layer. The next version projected the same offset into three decoder resolutions—H, $H/2$, and $H/4$—so that coarse structure and finer spatial detail did not have to pass through a single insertion point. A deeper decoder and this direct multi-depth route produced sharper dense reconstructions than the single insertion point, but sparse fields and exact values were still not recovered reliably. We also compared concatenating the latent with RGB at the decoder entrance against letting the decoder process RGB first and injecting the latent at a later layer. The entrance version exposed every subsequent block to the perturbation; the later version protected the early RGB computation but left fewer blocks

to interpret the added field. Neither placement removed the underlying tradeoff.

The jump-link experiments gave the decoder direct access to additional detail at several scales. The encoder returned features at four resolutions, H, $H/2$, $H/4$, and $H/8$. Learned 1×1 projections added these features to the matching decoder stages. We also tested dropping the offset or the jump features on a subset of training batches so that the decoder could not depend completely on either route. These links gave the decoder high-resolution evidence and could improve the training round trip when those encoder features were present. They did not make the compact latent carry the same information. A latent generated without the encoder-side links could therefore fail to reproduce the reconstruction seen during training. The skip path had bypassed the very representation that later generation depended on. Thus a better round trip did not necessarily mean that the latent itself had learned a portable representation.

Finally, we changed who was responsible for recovering the offset. Direct injection supplied the encoder's offset to the decoder. A decoder-prediction variant tried to infer that offset from an early decoder feature and inject the prediction into later layers. A fixed-ratio variant instead derived an offset algebraically from the adjusted latent. The learned predictor did not recover the encoder's offset accurately enough for later decoder stages, and the fixed-ratio rule preserved the numerical relation without restoring the missing field. Nano, tiny, and small encoder–decoder variants, LCPE, offset dropout, and a complete-restoration attempt explored capacity, position, and robustness around these routes. The complete-restoration variant did not recover the full RGB and additional-channel output, whereas returning to zero-initialized output layers and direct injection made the reconstruction loss fall much faster. Taken together, these variants showed exactly what "injection location" changed: early injection perturbed the shared RGB computation, late injection reduced the remaining capacity for the new field, and jump links could bypass the latent without making it generatively self-contained.

Output range formed a separate experiment chain. WSG used a missing-value code alongside physical values whose natural ranges differed by channel. We first applied fixed per-channel normalization and explicitly restored the missing code after decoding. A plain unconstrained output could leave the legal interval, while clamp produced zero gradient outside its boundary. Sigmoid stabilized one experiment; hard sigmoid and a learnable range tested whether stronger boundary behavior or a wider interval would preserve rare values. Extending the range made zeros easier to produce, but after nine hours the reconstruction still contained only a rough face and part of a body. Range handling could prevent invalid numbers without creating the missing structure.

The loss variants asked a different question. Full-image MSE was dominated by the missing background, so masked MSE computed error only where the target held a valid value. This increased the gradient from rare foreground points but did not teach the model where an omitted point should appear. SSIM added local structure, and a diversity term discouraged a constant output; both changed the appearance of the reconstruction without reliably preserving exact sparse values. These trials separated numerical validity, local image similarity, and exact-value fidelity as three different objectives.

Finally, categorical decoding replaced continuous values with explicit bins. The delayed CE–MSE variant first learned the coarse continuous field and then activated classification over values collected from the data. It preserved discrete values more explicitly than regression, yet the underlying image boundary still allowed a small location error to delete a point. The lesson was concrete: changing the activation or loss could improve one failure mode, but no output objective could recover a distinction that the chosen boundary did not represent robustly.

These experiments produced a clear tradeoff. When the new channels were allowed to change the shared latent space strongly, RGB reconstruction and the original pretrained distribution changed with them. When the RGB path was protected, the new channels had too little freedom: their outputs became averaged, fine details disappeared, and sparse structures were especially difficult to recover. Adding another type of information could also require changing a latent space that had already been learned for all previous types. This stage gave us the first important lesson. WSG should provide a common place to store, address, and exchange world information. The model can still use the encoding best suited to each type of information. RGB, dense physical fields, sparse skeletons, audio, and text have different structures and levels of abstraction. A common interface lets these variables work together while their internal representations remain suited to their own structure.

14.2. Early Sparse Codecs Exposed a Representation Problem

Sparse information was the first hard case encountered during codec development. A DWPose frame contains only a small set of valid points, yet each point carries a person identity, keypoint identity, location, and confidence. The ordinary CNN VAE learned the empty background much faster than these values; a one-pixel error could delete a joint, exchange two identities, or assign a point to the wrong person.

We first changed the codec while keeping the sparse field on its full image canvas. This preserved the original pixel address of every point, but almost all loss came from learn-

ing the missing-value background. Masked reconstruction restricted the loss to valid pixels; it increased the relative signal from the points, yet gave the model little reason to predict where a point should exist when that location was missing from its output.

The next variants removed the canvas. A patch/MLP VAE packed only valid values into tokens, and a coordinate-augmented version appended each value's spatial address. Removing empty pixels reduced the imbalance and allowed the MLP to begin learning. It also destroyed the regular neighborhood structure that a convolution obtained for free from the image. Adding coordinates restored absolute location but not the full two-dimensional organization or the relation among people and joints. We had therefore exchanged one problem—too much empty space—for another—a list whose geometry had to be relearned.

We then changed the output type. Continuous sigmoid, hard-sigmoid, clamp, and range-learning variants still averaged nearby IDs and confidences. Per-channel classification heads instead predicted the actual integer bins with cross-entropy. A delayed CE–MSE objective first learned a continuous image and then activated classification over the observed values. These variants kept symbols sharper than regression, but a one-bin error could still change a joint identity, and a small spatial error could still delete the point entirely. This separated two requirements that the original "sparse image" had mixed together: a representation must preserve the address of a point and preserve its discrete identity. Fixing only one did not make the round trip reliable.

At this point we did not yet have TunnelingNet, categorical Heads, or the later heatmap boundary. The immediate consequence was more basic: a single shared codec could not make heterogeneous fields mutually compatible merely by increasing its capacity. We therefore kept the learned representations separate and began to study whether small networks could bridge them.

14.3. We Then Built Small Bridges between Latent Spaces

After the codec experiments, we kept the representations separate and tried to connect them with small learned bridges. The plan was attractive. A strong text-to-RGB model could remain at the center. A bridge could map RGB to depth, depth to RGB, normal to edge, or any other requested pair. If every bridge were small, thousands of conversions could be added around one strong model. This was our proposed escape from the shared-codec problem: adding one task would train one bridge while leaving the representation used by every old task unchanged.

We tested this idea with CNN and MLP bridges, rectified-flow bridges [27], repeated refinement, minibatch optimal-transport pairing, VQ codebooks, index prediction, Transformer sequence models [71], conditional prior–posterior models, KL regularization, best-of-k sampling, and joint-embedding objectives. These experiments covered continuous and discrete latents, local and global networks, direct prediction and iterative generation, and deterministic and distributional losses. We trained dozens of bridge variants; Table 18 places them beside the codec, sparse-structure, object, and scaling experiments that motivated them.

Because the paired variables were spatially aligned, we first expected a small convolutional translator to learn their local correspondence. The resulting bridge used two independent full-resolution CNNs, one for each direction, with roughly 3M parameters per path. We added residual blocks, attention, RGB image conditioning, per-channel whitening, and learnable alignment. This model could be trained, but a single forward pass had to solve the complete cross-space mapping. Increasing its size improved the training loss much more than it improved decoded quality. The main difficulty therefore lay in the relation between the two spaces.

We also tested what the bridge should optimize. Latent MSE plateaued in the high-dimensional spaces, while latent MAE gave a more useful training signal. Decoded pixel loss checked whether a numerically close latent retained the requested field after passing through its VAE. A random 4×4 latent mask increased the amount of context the bridge had to reconstruct, and cached FP16/FP32 decoding checks ruled out storage precision as the cause of blurred outputs. These controls exposed a recurring gap: a small latent loss could coexist with a visibly smooth decoded field. Numerical proximity in the cached space was therefore an incomplete measure of successful translation.

At this point we began to describe the problem as translation between two encrypted languages. Each VAE code is meaningful to the decoder that learned it, but its individual numbers do not have an independently readable meaning. Two paired codes can describe the same scene while using unrelated internal "vocabularies." This is why concatenating, adding, interpolating, or matching their global distributions did not automatically create a shared language. It also explains why high-dimensional MSE could keep decreasing while decoded images barely changed: a numerically closer code was not necessarily a more correct sentence in the target decoder's language.

We next asked whether the mapping was difficult because one forward pass had to finish the entire conversion. We replaced it with a small U-Net and rectified-flow updates so that the model could predict several corrections and repair its own latent estimate step by step. This improved the formulation of the path, but training through the complete sequence stored the activations from every correction, which required much more memory. In one such experiment, the outputs later became black. Sampling one

random flow time reduced the memory cost, while training then supervised an individual time point and inference accumulated errors along a complete trajectory. We consequently tested reflow, explicit intermediate supervision, repeated correction blocks, endpoint losses, random-time training, and noise-conditioned paths.

Late-stage instability suggested that many reference paths crossed or approached the same region with conflicting velocities. One branch therefore addressed path crossings before the velocity network saw them. For each minibatch, we computed the pairwise distance between sampled noise and target latents and used optimal-transport pairing to minimize the total starting distance. This produced shorter and less crossed reference segments. It made the training geometry cleaner, but it could not supply the semantic relation missing between two independently learned latent spaces. Reflow and path correction had the complementary behavior: they could straighten trajectories produced by a teacher, yet their quality was bounded by the teacher paths and their repeated integration cost. This distinction was important. A better path optimizer can improve transport after the two endpoints have acquired compatible meaning; it cannot create that meaning by itself.

We tested this distinction again in a rectified auto-coder (RAC). A frozen TAESD [72] supplied a known image latent, and the learned encoder–decoder followed a one-step or four-step path to and from that latent. We compared a fixed and a randomized time grid, intermediate-state supervision, and noise added to path states to encourage correction after a perturbed input. Path projections made curvature and oscillation visible. Four-step training and intermediate supervision constrained the trajectory, but did not by themselves produce a large reconstruction advantage. More importantly, an early normalization change made RAC latents incompatible with an existing PerFlow generator [22] even when RAC's own round trip looked reasonable. Restoring the native TAESD latent contract restored the correct interface. The experiment showed that path regularity, correction robustness, and compatibility with a pretrained target space are separate properties.

This question produced a separate family of experiments that we called the *shadow-variable route*. Its motivation was simple: keep the mature RGB latent as an anchor and store depth in a nearby code, so that the existing RGB generator could continue to operate while the new code carried additional world information. We first trained a direct RGBD RAC with forward decoding and reverse encoding. We then represented depth as a three-channel color map and passed both RGB and visualized depth through the official image codec, so that the two sides had the same channel type. A strict bidirectional version forced the RGB and RGBD paths to use matched spatial positions. Finally, the perturbed-RAC version wrote the new state explicitly as $z_{\text{shadow}} = z_{\text{RGB}} + \alpha\Delta z$, with the frozen native RGB latent providing the anchor and the learned perturbation carrying the extra information.

These variants revealed the same underlying behavior from different angles. Early in training, a native RGB latent retained some ability to complete the paired depth field. After roughly 57,000 updates, the direct RGBD RAC could encode and decode its own state almost perfectly, yet depth completion from the unchanged RGB latent had become worse. The encoder and decoder had agreed on a private RGBD coordinate system that lay near the RGB latent numerically but meant something different. Depth visualization changed the token type without removing this semantic gap, and additive perturbation made the gap explicit: small Euclidean displacement did not guarantee compatible coordinates. We therefore called the result a *shadow latent*. Like two allotropes built from the same elements, the native and shadow codes could look similar while supporting different transformations. A successful private round trip proved that its own encoder and decoder agreed; it did not prove that the code remained interchangeable with the pretrained RGB latent.

Next, we converted bridge prediction into classification over a VQ codebook. The hope was that selecting a valid code would keep the prediction close to a known target latent. In practice, the target-channel distribution used only a small part of a 2,048-entry codebook. One batch used 251 RGB codes (12.3%) and only 24 target-channel codes (1.2%). Allowing the codebook to update improved coverage but also moved the class labels during CE training. Freezing it stabilized the labels and preserved many unused codes. We tested smaller codebooks, warm-up updates, per-position 1×1 prediction, MLP/index-MSE objectives, and explicit utilization tracking. These changes made more codes active or stabilized their labels, depending on the variant, while decoded samples still carried the wrong cross-variable content. Codebook health and cross-modal translation were therefore separate problems.

Local networks might have missed long-range correspondence, so we tested a Transformer bridge to provide global context, followed by learnable invertible 1×1 alignment and explicit mean, variance, and covariance matching. The alignment losses made the two populations look closer in their means, variances, and covariances, while the target decoder still produced incorrect content at individual positions. Matching two clouds of codes did not tell the bridge which target code belonged to each source sample. We then asked whether the remaining gap came from one-to-many uncertainty. Later experiments with a CVAE-style [73] conditional prior and posterior, encoder KL regularization, best-of-k sampling, target-noised training, reverse-noise training, style latents, and conditional diffusion [74] showed the same practical pattern. The condi-

tional prior–posterior branch used the condition and target during training to estimate a Gaussian posterior, used the condition alone to estimate the inference prior, and minimized their KL divergence. It supplied an explicit random variable, while the posterior, prior, and bridge still had to learn a high-dimensional cross-modal decoder from the same paired latent samples. Pure velocity-field prediction improved the RGB direction, but this underlying burden remained.

The two directions also behaved differently. RGB-to-depth, RGB-to-normal, and similar estimation directions map a rich observation to a more constrained field. Their targets occupy a narrower set and usually became recognizable earlier. The reverse direction must reconstruct appearance that the geometric condition never specified: color, texture, illumination, background, and many small details can all vary while depth or normal remains unchanged. Direct regression averaged these alternatives. Adding noise created multiple paths, while a small bridge still lacked the visual prior needed to turn those paths into realistic RGB. In the VAE experiments the same asymmetry appeared as resistance of the mature RGB latent: a weak update barely moved its decoded appearance, while a strong update easily left the pretrained RGB manifold.

The same phenomenon appeared across many of these variants. The bridge loss could decrease while decoded quality changed much more slowly. In some VQ experiments, only a small portion of the codebook was used. Larger sequence models learned the training pairs but retained the mismatch at decoding. Distribution losses made the latent statistics closer, yet the target decoder still assigned its own meaning to each coordinate.

This observation changed our view of the problem. A latent vector is meaningful because of the generator or decoder that learned it. Two spaces may have the same size and similar distributions while describing different information. A bridge that reaches a nearby point numerically can still miss the set of outputs that the target generator knows how to produce. We therefore stopped treating conversion as a separate operation between two finished latent spaces. The conversion needed to follow the target generation process from beginning to end, so that the target prior could guide every step.

14.4. We Tested Conversion inside Pretrained Video Generators

The bridge experiments showed that a small network could improve path geometry without learning the missing target domain. We therefore moved the conversion into pretrained image and video generators, whose denoising process already contained a strong prior over the requested target. This stage preceded the final FLUX implementation [26, 31] and tested whether several modalities or targets could be learned together inside one generator.

The LTX branch [33] separated two ways of using a pretrained video generator. Deep feature injection kept the backbone fixed and added condition features at several internal depths. It asked whether external evidence could directly push an already learned video trajectory. The condition changed the motion, but the generated person became increasingly distorted and the video no longer preserved the backbone's normal appearance quality. Inserting a feature at several layers therefore did not guarantee that the frozen blocks interpreted it correctly. The native LTX LoRA [12] variant instead changed small weight updates inside the generator itself. Its loss decreased normally, but the requested modality changed the video only weakly even after we added modality tags and monitored their use. These two failures were different: deep injection supplied evidence in an unreadable form, whereas the LoRA route could modify the prior but had not yet learned a strong target-directed change.

The Wan [32] and UniVidX [39] branches then tested whether several visual variables could be generated as fixed video lanes. We first reproduced the UniVidX baseline, then added a WSG adapter and an official-style control route. A smaller Wan 1.3B path let us repeat the same logic while measuring memory and interaction among lanes. These models could learn a predetermined set of outputs together, but adding a lane changed the joint tensor, sampling balance, losses, and attention seen by every existing lane. The cost was therefore not merely one small new output head; it changed the training problem for the whole model.

Reference conditioning provided a second comparison. Fun InP inserted a reference through the generator's supported inpainting route and produced usable guided video. Our VACE [75] reference integration passed the reference through a different control mechanism and produced incorrect video, so that branch was stopped. The contrast showed that "providing the same reference" was not a complete method description: the route by which the generator read the reference determined whether its pretrained prior remained usable.

We next varied how the joint lanes were sampled. Native multi-lane training materialized all outputs and could exceed memory. Sample-one training chose one lane per update, which reduced that update's memory but still changed a shared parameter set, so learning or revising one route could alter the others. Splitting conditions into groups of roughly four reduced the simultaneous load but did not remove interference. Fusing several conditions into one representation lowered the loss without producing a clear condition-specific output. Merging them into a larger token sequence moved the memory cost into attention. Modality tags told the model which lane it was seeing, but did not remove the shared optimization pressure.

Multilayer residual control changed the unit of learning. Each condition–target relation received a residual path through several depths of the same frozen generator. A route could now be trained, loaded, or revised without changing the parameters of another route. This experiment did not merely save memory; it separated capability acquisition from capability composition. It supplied the direct precursor to TunnelingNet.

This stage supplied two concrete results. A pretrained target generator was a better place for conversion than a bridge between finished latent spaces. At the same time, dense joint generation did not provide a practical route to hundreds of independently installable relations. We next kept the target prior shared and made each condition–target path explicit.

14.5. We Moved the Conversion into the Generation Process

This led to a simpler formulation. For each task, we identify the information that is already known, call it the condition, and identify the information that must be produced, call it the target. The model keeps the condition clean and generates the target along a normal generative trajectory. The learned path only needs to tell the shared generator how the condition should redirect that trajectory.

Direction is important. RGB→Depth and Depth→RGB use the same two variables, but the first estimates a relatively constrained geometric field and the second generates many possible appearances. They have different uncertainty and require different evidence. We therefore define a visual task by its condition, target, and direction.

Once we used this definition, many familiar task names became cases of the same operation. RGB-to-property tasks cover estimation and understanding. A world property used to generate RGB gives conditional generation. One world property used to generate another gives cross-property control. RGB combined with a current observation supports reconstruction and inverse rendering. Text joins the same process as semantic evidence. These directional relations became the four interfaces used by WSM.

The original deduction was even more concrete. Let A denote noise, M the RGB image, and N additional world information. An ordinary image generator learns $A \to M$; a WSG generator learns $A \to (M, N)$. A rectified-flow model can also invert the observed image from M back toward A. Estimation can therefore be written as $M \to A \to (M, N)$: invert the known RGB observation, regenerate it, and reveal its associated world information during the same generation process. In this view, estimating N from RGB is conditional generation with RGB as the condition. Reversing the direction makes N the condition and RGB the target; supplying RGB together with one world field to recover another gives reconstruction or inverse rendering. The four interfaces follow directly from which information is known and which information is generated.

The earliest notes stated the four cases in plain terms: text generates RGB; RGB generates additional world information for estimation and understanding; additional world information generates RGB for conditional control; and RGB generates an organized structural representation for reconstruction. Later experiments expanded the same grammar to cross-property translation and mixed conditions, but did not change its basic form. This also gave us a testable expectation: a stronger shared generator should improve more than image synthesis, because estimation and control now reuse its generative knowledge through a common backbone.

This was the central insight behind the Unified Law:

> *Visual tasks can be organized as directed conditional generation among RGB, text, and additional world information.*

WSG gives these variables stable identities and spatial boundaries. The four interfaces describe their main directions. The law describes the shared task form across possible neural architectures.

The final mechanism emerged through several target-generator experiments. In the Wan/UniVidX branch, we first assigned a LoRA to a target modality and mixed several conditions while training that target. This verified that one target prior could receive different evidence, but it also exposed coupling: adding a condition changed the shared training distribution, and adding many target lanes changed memory and attention for the entire model. We then moved to FLUX/OminiControl [36] and made the directed relation explicit. Each experiment selected one input–target task and updated the lightweight parameters that redirected the target denoising path.

Several ablations determined how that evidence should enter the generator. The native target-first version placed the reference beside the noisy target in the generator's existing input organization. The condition-lane version encoded the same reference through a trainable lane and injected its features into the target stream. Both preserved recognizable evidence, while the trainable lane learned noticeably faster in the direct A/B comparison. Thus, keeping the reference content fixed did not make its placement interchangeable: the target stream had to receive that content in a form its attention blocks could use.

We next changed the interaction itself. The original single condition interaction failed to converge on one Edge route; a second condition-attention path made the same data and target train. Task-type switches then selected the interaction used for estimation, conditional control, inverse rendering, and narrow generation. The switch did not rename one generic computation: it changed whether evidence en-

tered as an observation to interpret, a structure to follow, or an incomplete state to reconstruct. These trials made direction an implemented property of the route and recorded it explicitly in the route definition.

Finally, a high-noise anchor supplied an early global target estimate and prevented initial trajectories from collapsing to black. In a separate freeze test, the condition-side LoRA remained fixed while the target-side intervention continued to learn useful control. The model therefore did not need to rewrite the condition representation for every new target; it needed a trainable path that translated fixed evidence into residual pushes along the target trajectory. Together, these tests turned the abstract condition–target relation into a concrete intervention at several depths of one pretrained generator.

We separately tested whether a generic change to the dense flow objective could improve every route. The three levers were a low-noise timestep shift, Min-SNR weighting, and Huber loss. Huber improved the early part of training; Min-SNR retained more benefit later; enabling all three slowly collapsed toward the mean. We therefore left the ordinary objective as the default and retained the levers for task-specific use. The result was useful precisely because it showed that one global loss recipe was weaker than selecting evidence and supervision according to the target.

A controlled task-agnostic comparison isolated the underlying flow objective. Across Normal and Edge targets, plain rectified flow, direct clean-latent prediction, a high-noise anchor, and multi-resolution noise obtained mean fixed-time validation errors of 0.6383, 0.4247, 0.3662, and 0.6211, respectively. The anchor was the strongest of these four. It made global target structure available early without changing the condition–target definition, and it became part of the later training recipe.

Fine articulated targets prompted a second objective study. We first compared gradient, focal [76], direct clean-latent, and foreground-IoU supervision. We then warm-started two endpoint objectives—one in latent space and one after decoding—from a model that already generated recognizable pose. Training an endpoint objective from the beginning produced black outputs; adding it after warm-up still erased fingers and other sparse details. Six further continuations tested topology, a low-noise detail anchor, soft distance, multi-scale gradient, gradient direction, and hard-foreground mining. None improved the result enough to replace the gradient-supervised baseline. These trials showed why a loss that is reasonable for a dense endpoint can still suppress the small structures that define a pose.

We also tried to acquire many relations in one update while keeping a separate LoRA for every target. To reuse one condition computation, we froze the condition-side LoRA and packed eight target LoRAs behind the same condition encoder. A full eight-target graph exceeded memory in target attention. Four- target chunks still failed when their graphs were retained until one backward pass. Immediate backward after each four-target chunk made the update fit, and two-target chunks finally supported the complete eight-LoRA experiment on one GPU. The working version improved aggregate throughput by roughly two times, not eight times: target attention, target-specific LoRA projections, activations, losses, and gradients still grew with the number of relations. These variants made packed LoRA a useful scheduler, but they failed our intended scaling criterion of acquiring many capabilities at close to the cost of one.

We then tested the complementary sharing rule: one target LoRA for many conditions. One Edge LoRA first alternated among eight condition modalities. We also split dense conditions into groups of about four, fused multiple conditions into one training route, and expanded the candidate pool toward all 38 conditional-control relations. The small pilots could lower their training loss, but groups of about four already showed interference, fused conditions were only weakly trainable, and the larger Edge experiment had usable paired intersections for only 11 of the 38 candidate conditions. Its early true inference remained unstructured while separately acquired Edge tunnels already formed recognizable outputs. More importantly, every new condition updated the same target parameters, so adding or repairing one capability could change all previous capabilities. Shared-target LoRA therefore also failed as the large-scale replacement: it provided neither the quality nor the independent installation and revision of the final TunnelingNets.

The final design kept the pretrained backbone shared and gave each directed contract a light multilayer residual path. Native Joint Attention (NJA) supplied the ordinary joint baseline; Modality Tunneling Attention (MTA) exposed the condition features to a small internal tunnel and returned residual pushes at multiple generator depths. Zero initialization began from the unchanged native trajectory, while multilayer control allowed the push to accumulate through the generation process. Together, the installed paths form one modular WSM, and a requested task activates only its corresponding TunnelingNet.

14.6. We Extended the Framework to New Target Structures

After the directed generator and its internal tunnels were established, we returned to the sparse-data problem that the early codecs had exposed. The goal was no longer to force every variable into one VAE. We instead designed a faithful boundary for each target and let the shared generator learn the condition–target relation behind that boundary.

For pose, the first change enlarged every isolated point into a Gaussian blob. The VAE could now see enough fore-

ground area to preserve a location, and an inverse codec recovered the point through peak detection and duplicate suppression. The first version stored person and keypoint identity in the blob's color. A small reconstruction error could shift that color into another valid identity, so the recovered point could remain in the right location while being connected to the wrong person or joint.

We next changed where identity lived. Person buckets separated people before decoding. Body-part partitions separated body, hand, and face points; a body-only version measured the easier subset; and a full-body codebook tested whether one discrete vocabulary could preserve all identities. These layouts reduced particular collisions, but every new partition added lanes and still left identity vulnerable when it was carried mainly by a generated value.

The fixed-slot design made identity part of the address instead. Each person–keypoint pair received a predetermined slot, while the generated heatmap in that slot carried only location and confidence. Pack3 placed three such heatmap lanes in one RGB-like image without mixing their identities. The generator no longer had to reproduce an exact semantic color to say which joint it meant. This was the first pose boundary in the sequence that approached a reliable round trip: stable structure carried identity, and continuous generation carried only the quantities it represented well.

Compact structures followed a parallel route. Sequence heads emitted point IDs, coordinates, boxes, skeleton nodes, body parameters, or audio tokens with task-native ordering and losses. We first tested this idea on human parameters with a pretrained ScoreHMR [77] structure prior. One variant cached a feature for the whole image; the other aligned the image feature to the person crop before predicting WSV pose and shape parameters. Both avoided rasterizing the parameter vector into a dense field, while crop alignment removed much of the irrelevant scene. The structured output became coherent much earlier than the rasterized alternatives. This was the first clear example where a target-native structural decoder was a better boundary than an image codec for the same visual evidence.

Object boxes made the distinction more concrete. The first structured version predicted center and size slots directly. It was then given spatial image features so that the sequence knew where evidence occurred, IoU matching so that predicted and reference objects could be paired without a fixed input order, and focal supervision so that the many empty slots did not dominate learning. Duplicate suppression and NMS converted repeated candidates into a usable set of boxes. In parallel, we trained MTA and NJA routes to draw every box as a filled pseudocolor region. Those dense alternatives had to reproduce both exact geometry and an arbitrary instance color at every pixel; their boundaries and instance identities were visibly worse, so we stopped them. The comparison showed why the structured route worked: center and size remained one atomic object record, whereas the raster route spread the same relation across many independently generated pixels.

General objects required topology as well as coordinates. A fixed joint-ID/XYZ layout worked for a known human skeleton but could not describe an object whose parts and connectivity changed from sample to sample. We replaced it with a breadth-first serialization: every node stored its parent slot, a relative offset from that parent, and an absolute coordinate for recovery and checking. Breadth-first order made the parent available before the child and gave a variable object graph a deterministic sequence boundary.

We then asked how such a tree should be initialized and refined from a point cloud. K-means proposed representative points, and an MST supplied a connected starting topology. A direct refiner predicted the final tree in one pass; a diffusion refiner treated the tree as a noisy structured target; and an anchor-based refiner preserved the deterministic proposal while learning local corrections. K-means plus MST already produced a connected geometric proposal; the learned refiners could still move nodes or edges into unstable structures and were not reliable enough to replace it. The useful result was the representation lesson: for variable objects, topology, traversal order, parent identity, and permutation are part of the target definition, not bookkeeping that can be reconstructed after coordinate generation.

Audio extended the same task form beyond image-like fields. The first route used AST [78] features to classify an audio source without generating sound. The second combined RGB and aligned audio to predict the box of the sounding object. After full-audio PCM normalization and frame-level alignment were corrected, the classifier separated the source classes in the training data and the localization route placed boxes on sounding regions. This showed that the graph could mix two codecs when their time indices and semantic roles were explicit.

We then tested generation in the opposite direction. One route predicted EnCodec [79] tokens from visual evidence; another prefixed MusicGen [80] with image features. Both asked the model to infer a waveform-level result from an image that did not specify pitch, timing, timbre, or background sound. Their outputs remained weak and were kept optional. The failure was not evidence that audio could not join the graph; it exposed the same directional uncertainty seen in geometry-to-RGB generation, now with even more information absent from the condition.

The final audio variants converted sound into spatial evidence. Dense audio-semantic masks tried to paint the sounding region, while an RGB-plus-audio selector chose a region already supported by the image. The selector was more reliable because audio only had to disambiguate visible candidates. Across the six routes, the important distinc-

tion was therefore not "image versus audio," but what the condition actually determined about the target.

Categorical targets came later and were studied on working TunnelingNet routes. The first pilots kept the dense flow target but changed how classes influenced training. A fixed-class gated route restricted the allowed labels, and a balanced-flow route increased the weight of rare classes. They made local denoising previews look plausible, while full generation still produced noisy boundaries and unstable class regions. This established that a preview made from a partly correct target was easier than generating the categorical field from noise.

Prototype variants assigned every class a legal target color and trained pixels toward those prototypes. A self-contained prototype head, plain prototype CE, and a fresh prototype-CE experiment were compared with two multi-label designs. One multi-label design predicted the set of classes in the image; the other used a predicted foreground mask before class recovery. When the direct color constraint was removed, both multi-label routes could collapse to background. Without a separate classifier, prototype generation often produced fragments of several legal colors and lost regional coherence. The outputs were categorically legal at individual pixels without being a correct segmentation.

We then compared balanced CE, a combined objective, a global-margin loss, prototype CE, and prototype CE with focal emphasis under end-to-end generation. Global-margin and prototype CE produced the strongest candidates; focal emphasis did not by itself recover exact identities. Across these variants, continuous generation was good at drawing spatial support and less reliable at choosing an exact symbol. This pattern appeared across the loss variants and motivated the later separation between generated geometry and symbolic class recovery.

We therefore separated spatial support from symbolic recovery. Per-Head and multi-condition refiners, foreground recovery, and coarse-label remapping used the generated region and task definition to restore discrete identities. This two-stage design retained the generated spatial support and corrected more pixels to legal task labels without asking the flow model to solve geometry and exact symbol recovery with one continuous value.

We finally tested variables outside the original numeric WSG channels. MP-100 [81] Universal Pose compared full-frame rendering, corrected orientation and letterboxing, instance crops, gradient supervision, and focal supervision. The full-frame route collapsed toward black; instance crops with gradient supervision produced the strongest result. From that baseline, endpoint, topology, detail-anchor, distance, multi-scale-gradient, gradient-direction, and foreground-mining continuations showed that dense endpoint improvement could erase the fingers and small structures that define the pose.

MAGICK transparency and IC-Light relighting then tested whether a new variable had to occupy a preassigned numeric WSG channel. For transparency, we introduced a named alpha target and trained an RGB-to-transparency route for 1,000 updates. For relighting, the condition paired RGB with an environment-light variable and the target remained RGB; after 2,800 updates, the relighting model gave recognizable changes under fixed test lights. These routes used the same condition–target machinery as numbered channels while keeping their own descriptive node names. The result separated the stable graph interface from the growing set of variables: a new world attribute could be added without renumbering the original channel standard.

The final runtime experiment tested whether separately acquired relations could behave as a system. In the first chain, the unmodified backbone generated RGB from text and an installed RGB-to-world TunnelingNet converted that image into a second field. The generic version accepted an intermediate node produced by one route and selected the next compatible route from the graph. Each stage loaded only its own tunnel, kept its own learned parameters and provenance, and passed the typed result forward. No joint retraining was used. This changed the meaning of the task graph into an executable set of available transformations that could be composed at run time.

The later extensions confirmed the principle that the earlier experiments had gradually revealed:

> *Share the generative prior, preserve each target's required distinctions, and express every visual task as a directed condition–target relation.*

14.7. The Experiments That Led to the Unified Law

Table 18 summarizes the main hypotheses in their development order, while keeping variants of the same idea together. The parenthetical count in each family reports the named methodological variants represented by that row. Within every family, the account follows the same sequence: the observation that motivated the test, the variants actually tried, what happened, and what we learned. Additional experiments changed network size, loss weight, schedule, representation, or integration depth. This is why the experimental study became much larger than a comparison among network blocks. Its repeated outcome was structural: faithful codes preserve the required distinctions; path geometry, distribution matching, and random variables solve different parts of optimization; and realistic decoding still requires knowledge of the target domain.

The family view is useful for following the overall chronology, but it can hide the particular variant that produced each lesson. Table 19 therefore opens the most consequential latent-bridge families into individual experiments. Each row isolates one hypothesis, the change made to test

it, the observed behavior, and the conclusion carried to the next experiment. These rows are a finer account of variants already included in the 198-variant total; they do not inflate that count.

Table 18. **Experiments behind the Unified Law.** The 60 families contain at least 198 explicitly named methodological variants. Each row states the motivation, variants tested, observed result, and concrete lesson carried forward.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
Expanded RGB VAE (1)	We expected the pretrained RGB latent to contain spare capacity for world information, so we enlarged its encoder and decoder boundaries for additional WSG channels.	Additional channels changed the latent distribution immediately; preserving RGB and learning precise new fields competed for the same limited code.	Simply widening a mature RGB codec was not a scalable solution: every new field altered the shared latent and risked degrading all fields already stored there.
Frozen RGB path (1)	Because the expanded codec disturbed the mature RGB prior, we froze its RGB path and trained new layers around the encoder and decoder.	RGB survived better, while the additional channels had weak influence and decoded toward smooth averages.	Freezing protected RGB but starved the new fields of influence. Preserving an old capability and giving a new capability sufficient representational freedom require separate mechanisms.
Progressive TinyVAE expansion (6)	We froze the pretrained RGB path, widened hidden layers in 16-channel stages, expanded the latent from four through eight channels, and trained each new stage on full 256-channel WSG samples.	Dense fields began to reconstruct, sparse pose did not, and processing the complete WSG tensor reduced throughput from the original TinyVAE by a large margin.	Progressive growth did not remove the mismatch among dense and sparse fields, and its full-tensor cost grew immediately. More codec capacity alone could not solve heterogeneity or scaling.
Offset-routing schemes D–F (7)	Starting from $z' = z_{\mathrm{RGB}} + \Delta z$, we compared single- and multi-depth direct offset injection, decoder-entrance versus later latent injection, four-scale encoder–decoder jump links, decoder-predicted and fixed-ratio offset recovery, deeper decoders, LCPE, dropout controls, and complete restoration.	Direct injection at several decoder depths and a deeper decoder helped most. Four-scale jump links could bypass the compact latent but transferred poorly to latent-only generation; decoder-predicted offsets were difficult to optimize, fixed-ratio recovery did not improve the result, and complete restoration failed.	The location changed what the model learned: early injection disturbed the shared RGB computation, late injection left less capacity for the new field, and jump links improved access to encoder evidence without making the latent self-contained. Initialization and information path mattered more than simply adding parameters.
LayerDiffuse-style auxiliary path (3)	We hypothesized that a separate lane could carry new fields without rewriting RGB, so we added an auxiliary encoder–decoder interaction modeled on transparency encoding.	It carried extra information and produced the strongest result in this family, but precise dense fields and sparse structures still exceeded the small protected channel.	A protected auxiliary lane worked for a small amount of added information, but its averaged outputs and limited bandwidth ruled it out as the carrier for a growing world representation.
Independent channel VAE (3)	Shared-codec interference suggested that each variable should first reconstruct in its natural space, so we trained channel-specific continuous VAEs.	Individual dense fields became easier to reconstruct, but every modality acquired its own coordinate system and distribution.	Separate codecs made individual fields reconstructable, but created mutually incompatible latent coordinates. This result directly motivated the later bridge experiments.
Sparse categorical VAE (4)	Continuous regression averaged away sparse IDs and points, motivating categorical decoding, integer bins, CE losses, and hybrid CE–MSE objectives.	Discrete identities were preserved better than under regression, yet convergence was sensitive and small errors removed points or changed IDs.	Replacing regression with categories preserved IDs better, yet tiny decoding errors still destroyed points. Sparse information required a boundary that preserved both symbols and locations.
Sparse patch/MLP VAE (3)	A dense CNN spent most supervision on the empty canvas, so we tested a patch/MLP codec, valid-value tokens, and coordinate-augmented tokens at a comparable parameter budget.	Removing missing pixels reduced the imbalance but also removed the image's implicit spatial organization; retaining the canvas restored the imbalance.	Removing empty pixels fixed class imbalance but erased the canvas coordinates; retaining them restored the imbalance. The next representation had to preserve addresses without learning mostly background.
Cached paired latents (1)	Repeated image I/O and encoding obscured the mapping study, so we cached paired RGB and channel latents and trained directly on them.	Bridge training became much faster and the paired-space problem could be studied directly.	Caching removed image I/O and codec cost from the experiment. The remaining failures came from cross-space translation.

Continued on the next page.

Table 18 continued.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
Latent objectives and masking (4)	We compared latent MSE and MAE, decoded pixel supervision, random 4×4 latent masking, and FP16/FP32 cache checks to determine whether optimization or storage caused blurred decoding.	MAE improved high-dimensional optimization and pixel loss exposed failures hidden by latent distance; cache precision was not the main cause.	A low latent error was not evidence of a correct target after decoding. From this point onward, we judged bridges through decoded outputs and complete trajectories.
Seven local interpolation rules (7)	We asked whether paired spaces were locally smooth despite global misalignment, testing nearest-neighbor, RBF, thin-plate, barycentric, local-linear, manifold, and neural interpolation.	Nearby RGB codes did not reliably identify nearby target codes; larger neighborhoods smoothed incompatible targets.	No local interpolation rule made neighborhoods in one latent space meaningful in another. We therefore stopped assuming that paired modalities shared a usable local geometry.
Direct "God Bridge" (1)	If the paired relation was sufficiently regular, a compact mapper could solve it in one pass; we therefore trained a direct bidirectional latent bridge.	The network trained, but its loss initially failed to decline meaningfully.	A compact endpoint mapper could not discover the full cross-space relation in one pass. This failure motivated explicit paths and iterative correction.
Attention bridge (3)	Direct CNNs might lack global context, so we added attention, varied bridge width, and reduced the model to about 4M parameters.	Global interaction did not remove the endpoint mismatch; parameter count was not the main bottleneck.	Global context and additional parameters did not repair incompatible endpoint meanings. The bottleneck was not receptive field or network width.
Bidirectional CNN bridge (2)	Spatial alignment suggested local correspondence, while the two directions could require different functions; we trained independent full-resolution residual CNNs for both directions.	Training loss decreased and coarse structure appeared, while decoded quality improved slowly and differed strongly by direction.	The two directions learned at different rates and qualities. This established direction as part of the task, not merely an inference-time reversal of one bridge.
Whitening and fixed normalization (1)	Scale mismatch might dominate bridge optimization, so we normalized both cached spaces with measured statistics and inverted the transform after prediction.	Scale imbalance and optimization instability decreased; semantic mismatch remained after the first two moments aligned.	Matching scale improved numerical optimization but left the semantic mapping unresolved. First- and second-order statistics were only conditioning tools, not cross-modal alignment.
Learnable invertible alignment (3)	We hypothesized that the spaces differed mainly by an invertible coordinate change, so we learned 1×1 transforms and matched their moments and covariances.	The populations became statistically closer, yet decoded targets could still be wrong at individual positions.	Even trainable moment and covariance alignment could make populations look similar while individual samples decoded incorrectly. Distribution alignment could not replace paired meaning.
Joint embedding and JEPA-style alignment [82, 83] (2)	Shared semantics might be learned before decoding, so we aligned paired observations and tested whether one representation could serve both sides.	Partial shared structure emerged, while decodable modality-specific information still required separate capacity and joint retraining.	Shared semantics helped organization but did not retain all target-specific information required for exact decoding. A common embedding was insufficient as the sole generative representation.
Multi-step RF repair (1)	One-step conversion appeared too large, motivating a small U-Net that could repair its latent estimate through repeated rectified-flow updates.	Backpropagation through all repair steps consumed large memory; one experiment later produced black outputs.	Iterative repair increased expressive power but made full-path training expensive and unstable. We needed a path formulation that could be supervised without storing every correction.
Random-time RF (1)	Full unrolled repair exceeded practical memory, so we sampled one flow time per update and reserved multi-step integration for inference.	Memory dropped sharply; local velocity errors accumulated over the full trajectory.	Random-time training solved the memory problem but exposed a train–test mismatch: small local velocity errors accumulated during full inference. Complete trajectories still had to be evaluated.

Continued on the next page.

Table 18 continued.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
Endpoint and intermediate losses (5)	Local velocity MSE could hide a poor final decode, so we added endpoint losses, intermediate supervision, alignment terms, and repeated refinement.	Endpoint metrics revealed failures hidden by velocity MSE, but full-sequence supervision again increased cost and sometimes conflicted with local flow fitting.	Endpoint supervision revealed failures hidden by local velocity loss, but unrolling restored the original cost and sometimes conflicted with path fitting. No single loss solved both locality and final decoding.
Minibatch optimal transport (1)	Path crossings and conflicting velocities might explain late instability, so we reassigned noise–target pairs to minimize total latent distance.	Reference segments became shorter and less crossed; semantic incompatibility between the endpoint spaces remained.	Optimal pairing shortened and uncrossed reference paths, yet the decoded target remained wrong. It improved path geometry but could not provide missing target knowledge.
Reflow and teacher-path correction (3)	A teacher could reach the target along unnecessarily curved paths, so we generated its trajectories and retrained on their endpoints to straighten them.	Better teacher paths improved later integration; weak or unstable teachers transferred their errors, and data generation added another training stage.	Reflow could straighten a useful teacher trajectory and could not rescue a weak teacher. We retained it as a path-refinement tool, not as the source of target semantics.
Noise ranking and noise conditioning (2)	Multiple outputs could share one condition, and random starts needed identifiable roles; we ranked and exposed noise while training diverse paths to a target family.	Paths diversified and stepwise learning accelerated, showing that random starts could carry useful variation.	Exposing the random start allowed one condition to produce distinguishable paths. It solved path identity and diversity, while leaving target realism to a stronger prior.
VQ bridge with CE (2)	Continuous regression could leave the valid target manifold, so we quantized latents and predicted code indices to constrain every output to a learned codebook.	In one batch, a 2,048-entry codebook used 251 RGB codes (12.3%) and 24 target codes (1.2%); frozen codes left many dead entries and updated codes moved the labels.	Quantization kept predictions on known codes but severe code imbalance and moving labels limited the bridge. Valid discrete symbols did not guarantee a meaningful cross-modal mapping.
VQ utilization variants (4)	Poor VQ results might come from dead or imbalanced codes, so we tested smaller codebooks, warm-up updates, utilization tracking, and alternative CE/MSE weights.	Utilization improved in individual experiments, while decoded cross-variable quality remained limited.	Improving code usage repaired the quantizer without repairing translation. This separated codebook health from semantic compatibility between the two spaces.
MLP index-MSE bridge (1)	If neighboring code IDs carried smooth structure, lightweight index regression could simplify prediction; we tested that hypothesis with an MLP.	The attempt failed because numerical distance between code IDs did not represent visual or physical similarity.	Code indices had no ordinal meaning, so index regression optimized an artificial distance. Discrete targets required categorical supervision or a learned metric.
Transformer bridge (1)	Local convolution might miss long-range cross-modal relations, so we replaced it with a global sequence model.	The Transformer also failed to close the decoded distribution gap.	A global sequence model still could not close the decoded gap. This ruled out the explanation that the bridge failed only because CNNs were too local.
Pure velocity field, CFG, and condition dropout (3)	Competing losses might obscure path learning and the condition might be ignored, so we isolated conditional velocity and added guidance and dropout controls.	RGB prediction improved because the model finally learned distinct paths, while the reverse and information-poor directions remained much harder.	A clean conditional velocity objective produced clearer paths, especially toward RGB, but reverse directions remained difficult. Path learning helped only when the target prior was adequate.
CVAE-style prior–posterior bridge (1)	Deterministic bridges averaged one-to-many outputs, so we learned a target-aware training posterior, a condition-only inference prior, and their KL alignment.	The explicit latent variable added stochasticity, but prior matching and realistic target decoding remained difficult; this strategy did not produce a useful bridge.	Adding a latent random variable represented uncertainty but also required learning a difficult target decoder and prior. Stochasticity alone could not supply target-domain knowledge.

Continued on the next page.

Table 18 continued.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
Best-of-k endpoint sampling (1)	One reference target could admit several valid paths, so we sampled multiple starts and optimized the candidate with the best endpoint.	It reduced the penalty for alternative valid paths but did not repair the decoder mismatch.	Selecting the best sampled endpoint tolerated multiple valid answers but did not improve the generator that produced them. Search could not substitute for a target prior.
Target-noised training (2)	Reversing a known corruption process might teach the target distribution more directly, so we noised target latents and learned conditional denoising and velocity.	It made the target distribution part of training and improved path diversity, while endpoint variance and decoded detail remained unstable.	Noising the target exposed its distribution during training and improved diversity, yet detail remained unstable. This pointed toward performing conversion inside a mature target generator.
Style-latent conditioning (1)	Geometry leaves color, texture, and illumination unspecified, so we added a compact latent to carry these missing appearance choices.	The new variable exposed the missing information explicitly, but learning its prior and its effect created another difficult bridge.	The style variable made missing appearance information explicit, but learning its prior recreated the original bridge problem. Geometry-to-RGB needed an existing appearance model.
Reverse-noise encoding (1)	A learned inverse might provide structured starting states for forward generation, so we mapped target latents back toward their noise representation.	The route provided an explicit inverse experiment but did not yield a sufficiently stable conditional generator.	A structured inverse start did not make forward generation reliable. Inversion and target generation were separate capabilities.
Conditional latent diffusion (2)	RF failure might be schedule-specific, so we replaced it with an established DDPM/DDIM-style [84] conditional latent generator.	Despite its stronger stochastic formulation, the implementation did not produce a useful bridge.	Changing from rectified flow to diffusion did not remove the failure. The decisive limitation was cross-modal target knowledge, not the generic noise schedule.
Rectified auto-coder paths (6)	We compared one-step and four-step RACs, fixed and random time grids, intermediate-state supervision, state-noise correction, and native-latent compatibility with an existing generator.	Multi-step constraints changed path geometry but gave no automatic reconstruction gain; altering the TAESD latent contract broke PerFlow compatibility until the native convention was restored.	Path smoothness, perturbation recovery, and compatibility with a pretrained latent were independent properties.
Shadow-variable and perturbed-latent routes (4)	We tested direct RGBD encoding, depth visualized through the official RGB codec, strict position-matched bidirectional encoding, and an additive perturbation around the frozen native RGB latent.	The private RGBD code achieved a nearly perfect round trip while losing completion from native RGB; visualization changed the representation but not the semantic mismatch; the additive route remained close numerically without becoming interchangeable.	A nearby code can become a stable private coordinate system. Distance from the native RGB latent, codec compatibility, and semantic compatibility must be tested separately.
Video-backbone branches (9)	Isolated bridges lacked target-domain knowledge, motivating LTX deep injection, native LTX LoRA, modality-tagged Wan LoRA, UniVidX baseline/adapter/control, a Wan 1.3B path, Fun InP [85] reference, and VACE reference integration.	Deep LTX injection and VACE reference produced unusable outputs; the LoRA routes trained but initially had weak effects; Fun InP was usable; UniVidX/Wan exposed temporal memory and lane coupling.	A pretrained video generator supplied the missing target prior, but only carefully placed interventions could redirect it. This moved the search from external bridges to internal control paths.
Joint video-lane strategies (7)	Joint generation promised shared computation, so we tested native multi-lane generation, sample-one training, split condition groups, fused conditions, merged-lane big tokens, multimodal lane tagging, and multilayer residual control.	Native multi-lane training could OOM; sample-one reduced per-step memory; groups of about four still interfered; fused conditions were barely trainable; residual control restored separately acquirable routes.	Joint lanes worked for a fixed small set and became memory-heavy and mutually coupled as the set grew. Large capability collections required separately installable paths.

Continued on the next page.

Table 18 continued.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
FLUX/Omini evidence placement (5)	We tested native target-first and trainable condition-lane references, dual condition attention, task-type attention switches, and a frozen condition-side LoRA.	Both reference modes preserved evidence, while the trainable lane learned faster in the direct A/B comparison; dual attention fixed an Edge non-convergence case.	Reference placement and attention topology changed convergence even with the same data and target. Evidence therefore had to enter according to the direction and uncertainty of each task.
Task-agnostic flow objectives (4)	We compared plain rectified flow, direct clean-latent prediction, a high-noise anchor, and multi-resolution noise on Normal and Edge.	Their mean fixed-time validation errors were 0.6383, 0.4247, 0.3662, and 0.6211; the high-noise anchor was strongest and prevented early global collapse.	The high-noise anchor improved global structure most consistently without changing the task definition. It became a training improvement, not a replacement for the condition–target grammar.
Dense flow objective levers (3)	We compared a low-noise timestep shift, Min-SNR weighting, and Huber loss, both separately and together, under deterministic validation.	Huber was strongest early, Min-SNR held up better later, and all three together slowly collapsed toward the mean; the ordinary objective remained the default.	Individually useful loss changes interfered when combined and eventually encouraged mean outputs. We kept the ordinary objective as the default and treated these levers as target-specific options.
Packed multi-target LoRA acquisition (5)	To share one condition computation while retaining one LoRA per target, we tested a frozen condition-side encoder, a full eight-target graph, four-target chunks with retained graphs, four-target chunks with immediate backward, and two-target single-GPU chunks.	Full packing and retained-graph chunking exceeded memory in target attention. Immediate chunkwise backward made eight LoRAs trainable, but aggregate throughput improved only about $2\times$, since target-side computation and gradients still scaled with the number of relations.	Packing was a useful scheduler, not a near-constant-cost capability expansion method. Sharing the backbone execution did not remove the target-specific activation, optimization, or parameter cost.
Shared-target multi-condition LoRA (4)	To compress routes with the same output, we alternated eight conditions through one Edge LoRA, split dense conditions into groups of about four, fused conditions into one route, and expanded toward the 38 conditional-control candidates.	Small pilots lowered loss, but four-condition groups interfered, fused conditions were weak, and only 11 candidate conditions had usable paired intersections in the larger experiment. Early generated outputs remained behind separately acquired Edge tunnels, while every condition modified the same weights.	A common target did not make its condition-to-target relations interchangeable. Large-scale capability growth required routes that could be acquired, installed, repaired, and removed without changing earlier capabilities.
NJA/MTA TunnelingNet routes (4)	We compared native joint attention, modality tunneling attention, zero-initialized residual control, and multilayer residual control inside the shared FLUX generator.	The internal tunnel redirected the denoising trajectory with a small active parameter set and could be trained separately for each input–output task.	A small internal residual path could redirect a strong shared generator without retraining its target prior. This was the first mechanism that combined low marginal cost with independent task installation.
Semantic heatmap and inverse codec (2)	Isolated pixels gave the VAE too little signal, so we expanded points into Gaussian heatmaps and inverted them with peak detection, NMS, and ID quantization.	Locations survived round-trip encoding more reliably, while semantic colors could still change and reconnect the wrong joints.	Enlarging points preserved their locations, but color errors still changed identity. A usable sparse boundary had to move identity out of fragile pixel values.
Person–keypoint slot heatmaps (2)	Color was carrying too many meanings, so we assigned every person–keypoint pair a fixed slot and packed three heatmap lanes into each RGB image.	The slot index stabilized identity and left the VAE to preserve only a smooth blob and its position; the representation approached a usable round trip.	Fixed slots made identity structural and left the generator to predict only location and confidence. This produced the first nearly reversible boundary for sparse pose.
Body-part and full-body heatmaps (3)	Collisions among people, hands, faces, and body points motivated body partitions, body-only maps, and a full-body codebook.	Partitioning reduced semantic collisions, although each extra partition increased the number of lanes and the caching cost.	Partitioning reduced collisions at the cost of more lanes. The experiment showed exactly when extra representation capacity was worthwhile: when it preserved a distinction needed for decoding.

Continued on the next page.

Table 18 continued.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
Direct sparse sequence heads (5)	Image decoding remained inefficient for compact outputs, so we predicted ordered slots for points, poses, boxes, skeleton parameters, and audio tokens.	Training became compact and task-native, while each structure required explicit ordering, masks, matching, decoding, and losses.	Typed sequences were efficient for compact outputs but required explicit ordering and matching. The unified task form could therefore support different codecs without forcing one tensor topology.
ScoreHMR WSV structure route (2)	We tested full-frame and crop-aligned cached image features in a pretrained human-regression model while predicting WSV-derived SMPL [86] pose and shape parameters.	The image-feature route learned quickly and provided the first successful direct structured adapter in this sequence.	A pretrained structural prior learned body parameters quickly from image evidence. Target-specific decoders could play the same role for sequences that image generators played for dense fields.
Atomic object-box slots (3)	We compared a structured center/size predictor with a Gaussian spatial prior against MTA and NJA routes that generated filled pseudocolor box masks.	The structured route became usable with spatial features, IoU matching, focal supervision, duplicate suppression, and NMS; both dense-mask routes were stopped after inferior geometry and instance colors.	The structured box route outperformed dense pseudocolor masks because it kept each object's coordinates together. Object membership had to be encoded as an atomic relation.
Object BFS tree (2)	Fixed joint-ID slots assumed a shared skeleton that general objects do not possess, so we serialized parent slots, relative offsets, and absolute coordinates in breadth-first order.	The target represented variable object topology more faithfully, although the resulting skeleton quality still required improvement.	Fixed human-style slots could not represent arbitrary object topology. Parent order and traversal had to become explicit parts of the target code.
Point-cloud skeleton initialization (4)	Pure tree prediction lacked a geometric anchor, so we derived k-means/MST-style initial skeletons from object point clouds and tested refinement, diffusion, and anchor corrections.	Initialization produced a promising geometric baseline; downstream refinement remained experimental.	Deterministic geometry supplied a useful skeleton before learning, whereas unconstrained refinement remained unstable. A strong analytical initialization reduced what the learned route had to discover.
Sparse and mixed audio routes (6)	We tested AST source classification, RGB-plus-audio source boxes, EnCodec-token regression, image-prefix MusicGen, dense audio-semantic masks, and an RGB-plus-audio region selector.	Source classification, localization, and the mixed region selector became usable; direct visual-to-audio generation remained weak and optional.	Aligned audio could classify or select a visual region, while direct waveform/token generation remained weak. Cross-codec tasks were feasible when timing and semantic roles were explicit.
Dense categorical-mask pilots (8)	Before introducing the refiner, we tested a gated fixed-class route, balanced flow, a self-contained prototype head, set-level and predicted-mask multi-label heads, pixelwise semantic prototype CE, and prototype CE with and without a multi-label classifier.	True generation overturned optimistic partial-denoising previews: balanced flow retained noisy geometry, multi-label variants collapsed to background when direct color constraints were removed, and classifier-free prototype CE fragmented masks across legal colors.	True generation exposed failures hidden by partially denoised previews. Spatial support and exact class identity were separate problems and had to be evaluated separately.
Categorical objective modes (5)	We compared class-balanced CE, a combined categorical objective, global-margin supervision, prototype CE, and prototype CE with focal regression on a second target.	Four matched categorical-objective experiments completed 1,000 updates; global-margin and prototype CE gave the strongest visual candidates, while focal emphasis did not resolve discrete identity by itself.	Global-margin and prototype objectives improved candidates without making continuous flow exactly categorical. Better losses narrowed the gap but did not remove the need for symbolic recovery.
Categorical region refiners (4)	Pseudo-RGB generation could draw the right support while crossing a class boundary, so we tested per-Head refinement, multi-condition refinement, foreground recovery, and coarse-label remapping.	Discrete accuracy improved for several Heads, with difficulty remaining for fine-grained classes.	A second stage could preserve generated geometry and recover exact labels. This became the practical solution when one continuous generator could not reliably satisfy both requirements.

Continued on the next page.

Table 18 continued.

Family (named variants)	Motivation and variants tested	Observed result	What this taught us
Universal Pose custom Head (5)	To add a category-independent articulated target, we compared full-frame rendering, letterboxing, instance crops, gradient supervision, and focal supervision on MP-100.	Full-frame training collapsed toward black; corrected geometry and instance crops helped, and the gradient-supervised RGB-to-pose pilot gave the strongest result.	Correct geometry, cropping, and gradient supervision mattered more than simply enlarging the model. This successful custom target also showed that new variables need not consume fixed WSG channel IDs.
Pose-detail objective pilots (8)	Starting from a recognizable gradient-supervised pose route, we tested latent and decoded endpoint losses, topology, a low-noise detail anchor, soft distance, multi-scale gradient, gradient direction, and hard-foreground mining.	Endpoint losses produced black outputs from scratch and eroded fingers after warm-start; the other six continuations were also weaker than the original gradient-supervised baseline.	Endpoint-oriented losses repeatedly removed fingers and other sparse details. We kept the gradient-supervised baseline and learned that dense endpoint improvement could oppose articulated fidelity.
Named custom-variable routes (2)	We added an RGB-to-transparency route from MAGICK data and an RGB-plus-environment-light to relighted-RGB route from IC-Light data.	The transparency route completed 1,000 updates; after 2,800 updates, the relighting model produced recognizable results under fixed test lights.	Transparency and relighting confirmed that named graph variables generalized beyond pose. The task graph could grow without redesigning the original numeric channel allocation.
Runtime route composition (2)	We tested text-to-RGB followed by an RGB-to-world route and a generic intermediate-node chain assembled from installed tunnels.	The stages executed without joint retraining and preserved separate provenance for every edge.	Installed routes could be chained without joint retraining while preserving their provenance. Independently acquired capabilities therefore formed an executable system.

14.7.1. Latent-Bridge Variants in Development Order

For this table, a *source latent* is the compressed tensor produced by the frozen source encoder, a *target latent* is the tensor expected by the frozen target decoder, and a *bridge* is the trainable mapping tested between them. Each row states exactly what changed relative to the preceding design, why we made that change, what we observed, and what that experiment allowed us to conclude.

Table 19. **Variant-level chronology of the latent-bridge search.** This table unpacks the bridge families in Table 18. Rows follow the approximate development sequence, including points where work moved between families; the first column preserves each recurring research line without rearranging that chronology.

Family	Variant and key change	Why we tried it	What happened	What we learned
Direct map	**Direct "God Bridge"** Feed the source latent to one compact CNN and regress the paired target latent in a single pass.	The most optimistic hypothesis was that a compact endpoint mapper could solve the paired relation in one pass.	The bridge was trainable, but its loss initially showed little meaningful decline and its outputs lacked fidelity.	A small one-pass network could reduce latent error, but it could not reproduce the target images faithfully. We therefore stopped treating the entire cross-space conversion as ordinary pointwise regression.
Direct map	**Attention-augmented CNN** Insert attention blocks into the compact CNN so every latent location can use distant locations.	Convolution might miss global relationships, so attention blocks were added while the bridge was kept lightweight.	Global interaction did not remove the endpoint mismatch.	Because attention did not solve the mismatch, the failure was not caused mainly by a limited receptive field. Global communication alone could not translate coordinates whose meanings differed across the two latent spaces.
Measurement	**Cached paired latents** Encode every paired RGB/target sample once, save both latent tensors, and train the bridge only from this cache.	Image loading and repeated VAE encoding obscured the mapping experiment, so paired RGB and target latents were precomputed once and streamed during bridge training.	Training reached roughly thousands of updates per hour and became much easier to inspect, while decoded failures remained.	After removing repeated encoding from the loop, training became fast while the same decoded errors remained. This ruled out the data pipeline as the main bottleneck and isolated latent-space translation as the problem.

Continued on the next page.

Table 19 continued.

Family	Variant and key change	Why we tried it	What happened	What we learned
Alignment	**Whitening** Measure each space's mean and scale, standardize both before the bridge, and undo target standardization before decoding.	Mean and scale mismatch might dominate optimization, so each latent population was standardized with measured statistics and inverted after prediction.	Numerical conditioning improved, while semantic errors remained after first- and second-order scale alignment.	Standardizing both latent populations made the optimizer behave better, but the same samples still decoded to the wrong content. Similar means and scales therefore did not imply corresponding latent meanings.
Direct map	**Wider residual CNN** Enlarge each directional residual CNN to roughly 3M parameters while keeping the paired data and objective fixed.	Insufficient bridge capacity was a simple explanation for the plateau, so each directional CNN was enlarged to roughly three million parameters.	Training loss improved more than decoded quality.	The larger mapper fit the paired coordinates better without producing comparably better images. The plateau was therefore not simply caused by too few bridge parameters.
Measurement	**FP16 versus FP32 cache** Store and decode the same cached latent pairs in FP16 and FP32, then compare the resulting images.	Blurred decoding might have come from quantization in the cached latent pairs, so the same pairs were checked and decoded under both cache precisions.	The precision change did not explain the large visual gap.	Both cache precisions produced the same large visual gap. Quantization during latent storage was therefore ruled out as the source of the bridge failure.
Measurement	**Latent MSE** Predict the target latent directly and minimize mean squared error at every latent coordinate.	The first direct bridge treated paired coordinates as ordinary regression targets and minimized squared latent error.	The loss decreased and then plateaued while decoded details remained visibly wrong.	MSE could decrease even when the target decoder produced visibly incorrect details. A small coordinate error in latent space was therefore not a reliable measure of correct target content.
Measurement	**Latent MAE** Keep the same direct mapper and paired latents, but replace MSE with mean absolute error.	High-dimensional MSE concentrated gradient on a subset of large residuals, so the bridge objective was switched to MAE.	Optimization became steadier and late residuals were less dominated by outliers, without closing the decoded gap.	MAE made optimization less sensitive to large residuals but did not improve the decoded endpoint enough. Better numerical conditioning alone did not solve the semantic mismatch between the spaces.
Measurement	**Decoded pixel supervision** Decode the predicted latent with the frozen target VAE and add an image-space error against the target image.	To test whether a close latent remained correct after decoding, the predicted target latent was passed through its frozen decoder and compared in image space.	Pixel supervision supplied a stronger visual signal, yet fine structure improved slowly and the bridge still produced smooth averages.	Comparing decoded images exposed errors hidden by latent loss and gave a more useful training signal. Its slow, averaged outputs also showed that supervision alone could not supply the target-domain prior missing from a small bridge.
Measurement	**Random latent masking** Hide random 4×4 source-latent regions during training and ask the mapper to predict the full target.	The direct mapper might have copied local coordinates without learning context, so random 4×4 latent regions were hidden during training.	The intervention did not produce a clear improvement in decoded quality.	Forcing the mapper to infer hidden neighborhoods did not improve decoded quality. The dominant error was therefore not simple local copying or insufficient spatial context.
Direct map	**Image-conditioned bridge** Give the mapper both the source latent and features from the original RGB image.	The latent pair alone might omit information visible in the source image, so RGB image evidence was added as an explicit bridge condition.	Extra evidence helped individual reconstructions but did not make the two completed latent spaces mutually decodable.	Adding the source image helped some individual predictions, but the target decoder still interpreted many outputs incorrectly. More source evidence reduced ambiguity without defining the target latent language.
RF path	**U-Net repair bridge** Replace the one-pass mapper with a multiscale U-Net that repeatedly updates its current target-latent estimate.	A multiscale U-Net was introduced so the bridge could repeatedly correct its latent estimate.	Repeated repair increased memory sharply, and one experiment later produced black outputs.	Repeated U-Net correction added capacity, memory use, and instability, yet it did not reliably recover the target. Iterative repair was therefore not a practical substitute for a correct target-space model.

Continued on the next page.

Table 19 continued.

Family	Variant and key change	Why we tried it	What happened	What we learned
RF path	**Fully unrolled RF** Run every rectified-flow integration step during training and backpropagate the final endpoint loss through all steps.	All rectified-flow corrections were differentiated through so the final endpoint could supervise the complete path.	Activation storage grew with the number of integration steps and exceeded the intended lightweight budget.	Supervising the final endpoint through every integration step was possible, but memory grew with the full trajectory. This established global path supervision as useful but too expensive to serve as the default training procedure.
RF path	**Random-time RF** Sample one time t per update and train the velocity at that point, which avoids storing the full integration path.	One random time point on the reference path was supervised per update.	Memory fell substantially, while small local velocity errors accumulated during complete inference.	Training one randomly sampled time step greatly reduced memory, but small velocity errors accumulated into a poor final image. Good local flow predictions therefore did not guarantee a correct full trajectory.
Direct map	**Independent directional CNNs** Use one CNN for RGB-to-channel and a different CNN for channel-to-RGB; do not share their weights.	One shared bidirectional rule assumed that reversal preserved difficulty, so RGB→channel and channel→RGB received separate networks.	The estimation direction improved faster; the reverse direction remained substantially harder.	RGB-to-structure learned much faster than structure-to-RGB under otherwise similar settings. Reversing condition and target therefore created a different task with different uncertainty and could not share one symmetric rule.
Alignment	**Learned 1×1 alignment** Place a learned invertible 1×1 channel projection between the two latent spaces before translation.	If the two spaces differed mainly by a channel-wise coordinate transform, a learned invertible projection could align them before translation.	Population statistics moved closer, but individual predictions still decoded incorrectly.	The projection aligned population statistics, yet paired samples still decoded to the wrong targets. A channel-wise invertible transform was therefore insufficient to recover sample-level correspondence.
Alignment	**Moment and covariance matching** Add losses that match the predicted and target latent means, variances, and covariance matrices.	To go beyond whitening, mean, variance, and covariance penalties were added between paired populations.	Distribution plots looked more similar while output details remained wrong.	The two latent clouds looked more alike after matching their statistics, while output details remained incorrect. Distribution-level similarity therefore did not guarantee that corresponding samples carried the same meaning.
Alignment	**Joint-embedding alignment** Encode the paired RGB and target observations into one shared embedding, then decode the required modality from it.	Paired RGB and target observations were pulled into a common representation before decoding.	Shared structure emerged, but exact modality-specific information required separate capacity and joint retraining.	The shared embedding captured common structure but discarded details needed only by one modality. A common space could organize modalities, yet faithful generation still required target-specific information and capacity.
Probabilistic	**Encoder KL regularization** Add a KL penalty that pushes the bridge bottleneck toward a smooth reference distribution.	The bridge bottleneck was regularized toward a smoother common distribution.	The latent became easier to sample, while modality-specific detail weakened when regularization dominated.	KL regularization made the bottleneck smoother and easier to sample, but strong regularization erased modality-specific details. Sampling convenience therefore had to be balanced against the information required for exact conversion.
RF path	**Noise ranking** Assign a persistent rank/identity to each sampled noise start and condition the path on that identity.	Random starts were assigned stable ranks so different paths for one condition could remain distinguishable.	The bridge produced more differentiated trajectories and stepwise training became faster.	Giving each random start a stable identity reduced averaging among different possible paths. This showed that part of the collapse came from indistinguishable trajectories, although path identity alone did not solve target generation.

Continued on the next page.

Table 19 continued.

Family	Variant and key change	Why we tried it	What happened	What we learned
RF path	**One-step endpoint loss** Take one numerical integration step during training and penalize its predicted endpoint against the target latent.	A cheap endpoint term was added to correct trajectory drift, initially using a single integration step.	The endpoint number could become small while full-step inference and decoded output remained poor.	A low one-step endpoint error could coexist with poor eight-step inference and poor decoded images. Endpoint loss was therefore trustworthy only when it used the same trajectory and representation as actual inference.
RF path	**Matched multi-step endpoint loss** Replace the one-step approximation with the same eight-step integrator used at inference.	The endpoint computation was changed to use the same eight-step integration schedule as inference.	Training and visualization became more comparable, at greater computational cost.	Using the inference-time integration schedule made the training endpoint and evaluated endpoint refer to the same object. The additional cost was necessary to interpret endpoint improvements correctly.
RF path	**Checkpointed endpoint integration** Keep eight-step endpoint training, but discard intermediate activations and recompute them during backpropagation.	Multi-step endpoint supervision was retained while activations were recomputed during backward propagation.	Peak memory grew much less, although training slowed.	Recomputing intermediate states kept multi-step endpoint supervision within memory, although each update became slower. This provided a workable memory–time trade-off without changing the underlying translation problem.
Measurement	**Full-step decoded metrics** Run the complete inference integrator, decode its final latent with the target VAE, and score that decoded output separately.	Complete inference was evaluated separately from the differentiable endpoint term and then decoded through the target VAE.	Low local or aligned-space error could coexist with much larger full-trajectory and visual error.	Several models looked good under local training loss but failed after full integration and decoding. We therefore retained complete decoded inference as the decisive bridge evaluation.
RF path	**Minibatch OT pairing** Within each minibatch, use optimal transport to reassign sampled noise starts to target latents before flow training.	Intersecting noise-to-target segments could assign conflicting velocities to nearby states, so noise samples were reassigned to targets by minibatch optimal transport.	Reference segments became shorter and less crossed, while semantic incompatibility between endpoint spaces remained.	OT shortened and untangled the reference segments, but the decoded targets remained wrong. It improved transport geometry while leaving the missing target-domain semantics unchanged.
RF path	**Noise conditioning** Expose the sampled starting noise to the velocity network at every prediction.	The sampled start itself was exposed to the velocity network as a condition.	Path diversity improved, especially toward RGB, while difficult reverse mappings remained.	Exposing the sampled start helped the network distinguish multiple trajectories and increased diversity. Reverse mappings still failed, showing that trajectory identity could not replace knowledge of how to generate the target modality.
RF path	**Reflow** Use a trained bridge to generate start–endpoint pairs, then train a second bridge on those generated paths.	A trained bridge was used to generate teacher endpoints, and a second model learned straighter paths between the observed starts and endpoints.	Useful teacher paths could be straightened; weak or unstable teacher outputs were inherited.	Reflow successfully straightened paths only when the first bridge already supplied useful endpoints. Errors in the teacher endpoint persisted in the second model.
RF path	**Intermediate-state supervision** Integrate to selected interior times and supervise both the intermediate states and final endpoint.	Errors could accumulate between randomly supervised time points, so intermediate integrated states were explicitly constrained.	Path behavior became more regular, without an automatic gain in final decoded reconstruction.	The additional constraints made intermediate states smoother and more regular, but the final decoded target did not improve automatically. Controlling the path interior therefore did not supply the missing endpoint semantics.

Continued on the next page.

Table 19 continued.

Family	Variant and key change	Why we tried it	What happened	What we learned
RF path	**Repeated correction blocks** Apply the same lightweight correction block several times, feeding each predicted state into the next update.	The same lightweight correction rule was applied several times to let later steps repair earlier errors.	Error correction improved locally but multiplied integration cost and could amplify bias across steps.	Repeated updates repaired some local errors but also repeated systematic bias and increased inference cost. Iteration helped only when the correction rule had been trained on the off-path states encountered during inference.
VQ bridge	**VQ bridge with CE** Quantize target latents into codebook IDs and train the bridge as a per-location classifier with cross-entropy.	Continuous predictions could leave the valid target manifold, so target latents were quantized and the bridge predicted code indices.	Predictions remained on codebook entries, but target usage was extremely imbalanced and many codes were inactive.	Quantization kept predictions on valid codebook entries, but most target codes were rarely or never used. Discreteness prevented invalid coordinates without making the codes balanced or semantically predictable from the source.
VQ bridge	**Frozen VQ codebook** Pretrain the VQ codebook, freeze its entries, and train only the bridge that predicts their IDs.	The codebook was frozen to make CE labels stable throughout bridge training.	Labels stopped moving, but dead entries and poor target coverage remained.	Freezing kept class labels stable, but it also preserved dead codes and poor coverage of target patterns. Stable labels alone could not make an unsuitable codebook represent the target distribution.
VQ bridge	**Updated VQ codebook** Allow the VQ entries to update jointly while the bridge learns to predict their current IDs.	The codebook was allowed to adapt so more target patterns could receive codes.	Utilization improved while the class meanings moved during classifier training.	Updating activated more codes, but their meanings moved while the bridge classifier was learning them. Better coverage therefore introduced a moving-target problem for cross-modal prediction.
VQ bridge	**Smaller VQ dictionaries** Reduce the original 2,048-entry dictionary and repeat the same VQ bridge training.	The 2,048-entry codebook was reduced to test whether excessive capacity caused dead codes.	Utilization rose in some experiments, while decoded translation remained limited.	Reducing the dictionary increased code usage but did not recover accurate decoded translation. Dead-code reduction was therefore a codebook diagnostic, not evidence that the bridge had learned the target mapping.
VQ bridge	**Index-MSE MLP** Output one real number per location and regress the numerical VQ index with MSE before rounding to an ID.	Code IDs were regressed as numbers under the hypothesis that nearby indices represented nearby content.	The attempt failed because code indices had no ordinal geometry.	Regression failed because code IDs were unordered labels. A VQ bridge had to predict them categorically or learn a meaningful code metric.
Direct map	**Transformer bridge** Flatten latent locations into a sequence and replace the convolutional mapper with a Transformer.	A sequence model replaced local convolution to test whether long-range interaction was the missing ingredient.	The Transformer also retained the decoded distribution gap.	The Transformer added global context and more flexible interactions, yet the same decoded gap remained. The failure was therefore not specific to convolution or short-range processing.
RAC	**One-step RAC** Map the input directly to a TAESD-compatible latent with one learned rectified-auto-coder update.	A rectified auto-coder tested whether a direct learned path could reproduce a known TAESD latent contract.	It provided a simple baseline but little opportunity to correct perturbations along the path.	One-step reconstruction measured whether the endpoint could be reached directly, but it revealed nothing about recovery from errors along a trajectory. We therefore separated endpoint fidelity from path robustness.

Continued on the next page.

Table 19 continued.

Family	Variant and key change	Why we tried it	What happened	What we learned
RAC	**Four-step RAC** Replace the single update with four sequential learned updates and inspect every intermediate state.	The same auto-coder was expanded to four learned stages.	Path geometry changed and intermediate behavior became visible, without an automatic large reconstruction gain.	Four stages exposed and reshaped the intermediate path, but additional stages did not automatically improve the final reconstruction. Step count was a control variable, not a substitute for a compatible latent contract.
RAC	**Fixed versus random time grids** Train otherwise matched RAC models using either fixed integration times or newly sampled times per update.	RAC compared a prescribed integration grid with randomly sampled training times.	The two schedules emphasized different parts of the trajectory and neither alone solved compatibility.	The schedules changed which trajectory regions were learned best, while both retained compatibility errors. Time sampling affected optimization coverage but could not repair a mismatched latent interface.
RAC	**State-noise correction** Add noise to intermediate RAC states during training and require later updates to recover the clean trajectory.	Noise was added to intermediate RAC states so later updates learned to recover from off-path inputs.	Perturbation recovery improved while clean reconstruction changed less.	Training on perturbed intermediate states improved recovery when inference drifted away from the ideal path. Clean reconstruction changed little, confirming that robustness to accumulated errors had to be trained separately.
Shadow variable	**Direct RGBD shadow code** Train forward RGB-latent-to-RGBD decoding together with reverse RGBD encoding, then test both the learned code and the unchanged native RGB latent.	We expected reverse encoding to place RGBD around the mature RGB manifold while preserving the original RGB coordinates.	After roughly 57,000 updates, the learned RGBD code reconstructed almost perfectly, while depth completion from unchanged RGB became worse than early in training.	Encoder and decoder had formed a private coordinate system. Their agreement did not preserve the meaning of the native RGB latent.
Shadow variable	**Visualized-depth shadow code** Convert depth into a three-channel color image and encode it with the official image codec before learning the RGB–depth interaction.	Matching the channel type and codec was intended to remove the possibility that raw scalar depth was simply incompatible with an RGB encoder.	The representation became easier for the image codec to process, yet the cross-modal semantic gap remained.	Making a modality look like RGB solves a representation-format problem, but it does not make its latent coordinates mean the same thing as native RGB coordinates.
Shadow variable	**Strict bidirectional alignment** Replace the loose expand–pool projection with a subpixel-seeded mapping that gives forward decoding and reverse encoding matched spatial positions.	We tested whether the shadow code drifted because the two directions were allowed to use different spatial correspondences.	Position matching improved bookkeeping and exposed intermediate alignment error, but private-code reconstruction still did not imply native-code completion.	Spatial correspondence is necessary for a reversible interface, yet coordinate alignment alone cannot impose semantic identity between modalities.
Shadow variable	**Native-RGB additive perturbation** Parameterize the new code as $z_{\mathrm{RGB}} + \alpha\Delta z$, freeze the native RGB anchor, and restrict the learned contribution to a scaled perturbation.	This was the strongest test of the shadow hypothesis: extra world information had to remain close to a valid pretrained RGB code.	The route kept the new code close in magnitude while exposing conflicting objectives between RGB preservation and RGBD recovery.	Euclidean proximity can be enforced explicitly, but semantic compatibility cannot. A perturbation becomes useful only when the downstream generator has learned how to interpret the information carried by that perturbation.
RAC	**Native-latent compatibility** Remove the added latent normalization and use the original TAESD scaling expected by PerFlow.	An extra latent normalization was removed after it broke decoding by an existing PerFlow generator.	Restoring the original TAESD convention restored the interface.	The normalized codec could reconstruct its own samples while failing when connected to the existing PerFlow generator. Restoring native TAESD conventions showed that a useful latent contract must remain compatible with the pretrained model that consumes it.

Continued on the next page.

Table 19 continued.

Family	Variant and key change	Why we tried it	What happened	What we learned
Probabilistic	**Pure conditional velocity** Remove auxiliary alignment and endpoint objectives, leaving only conditional velocity prediction.	Auxiliary objectives were removed so the bridge learned only the conditional vector field.	RGB-directed prediction improved, while information-poor reverse directions stayed difficult.	Removing auxiliary losses clarified the flow and improved prediction toward RGB, but reverse directions remained difficult. The remaining gap followed the information and generative knowledge required by the target, not objective complexity.
Probabilistic	**Classifier-free guidance** Randomly drop the condition during training, then combine conditional and unconditional predictions at inference.	Condition dropout and guided inference were added to strengthen use of the source modality.	Conditional influence became easier to control; it did not repair the target decoder mismatch.	Guidance strengthened or weakened how much the model used an available condition. It could not create appearance or target-domain knowledge that the bridge had never learned.
Probabilistic	**CVAE prior–posterior** Train a target-aware posterior and a condition-only prior, and align them with a KL loss.	A target-aware posterior and condition-only prior were trained with KL alignment to represent one-to-many outputs.	The random variable represented uncertainty, but prior matching and realistic decoding remained difficult.	The latent variable represented multiple possible outputs, but the condition-only prior still could not decode them realistically. Modeling uncertainty helped only after the model had learned a capable target generator.
Probabilistic	**Best-of-k sampling** Generate k candidate starts or paths and backpropagate through the candidate whose endpoint is closest to the reference.	Several candidate starts were generated and the endpoint closest to the reference was selected for optimization.	Alternative valid paths were penalized less, but the generated candidates did not become intrinsically better.	Choosing the closest of several samples stopped penalizing every alternative as wrong, but it did not improve the samples being proposed. Selection handled evaluation ambiguity without strengthening the generator.
Probabilistic	**Target-noised training** Add noise directly to target latents and train the model to recover them while conditioning on the source.	Target latents were corrupted and conditionally denoised so training directly visited the target distribution.	Diversity improved, while endpoint detail and stability remained insufficient.	Denoising corrupted target latents exposed the model directly to the target distribution and improved diversity. Fine detail and stability still lagged, so exposure alone did not replace a mature target generative prior.
Probabilistic	**Style-latent conditioning** Add a compact latent carrying color, texture, and illumination alongside the geometric condition.	Geometry leaves appearance unspecified, so a compact latent was added for color, texture, illumination, and other residual choices.	The missing variables became explicit, but their prior and decoder formed another difficult bridge.	The style latent made the missing appearance variables explicit, but those variables still needed a learned prior and decoder. Merely naming color, texture, and illumination did not make them inferable from geometry.
Probabilistic	**Reverse-noise encoding** Train an inverse model from target latent to structured noise, then use that noise to initialize forward conditional generation.	Target latents were mapped backward toward structured noise to provide a better starting state for forward prediction.	The inverse route was measurable but did not yield a stable conditional generator.	The inverse model could move a known target latent toward structured noise, but the forward model still could not regenerate a stable target from that noise and the condition. Thus, learning an inverse route did not provide the target knowledge required for conditional generation.
Probabilistic	**Conditional latent diffusion** Replace rectified flow's velocity prediction along a continuous interpolation path with conditional DDPM/DDIM denoising across multiple noise levels.	Flow matching and diffusion define different learning problems. We tested whether iterative stochastic denoising could model the target distribution that the learned flow path could not reach.	The diffusion objective trained under these different dynamics, but its decoded targets were still not practically useful.	The same failure survived a change from velocity-field learning to stochastic denoising. This ruled out a flow-specific path or objective as the main cause and showed that the small bridge still lacked the target-domain knowledge needed to generate correct outputs.

14.8. How the Experiments Led to the Unified Law

The law did not begin as a slogan. It was the common answer left by a long sequence of failed and successful experiments. A single codec could reconstruct some world fields, but every newly added field changed the shared latent space. Small bridges could align statistics, shorten paths, or reconstruct their own private codes, yet they could not make independently trained modal spaces mean the same thing. Sparse points disappeared until we gave every point a visible address. RGB-to-depth learned more easily than depth-to-RGB because the second direction must invent appearance that depth never contains. Finally, conversion became reliable when the requested target was produced through a generator that already knew how that target should look. These observations gave us one simple description: known RGB, text, and world information are conditions; the requested information is the generated target; reversing the direction defines a different task.

WSM v1 was built after this experimental history and tests the resulting rule at scale. Its 1,148 evaluated tasks span 39 output groups and all four interfaces. Task-native metrics, qualitative matrices, external comparisons, and executable chains test many different condition–target directions under the same definition. The earlier experiments explain where the rule came from; the WSM study tests how broadly it works.

14.9. The Deeper Insight: a Model Must First Preserve What It Is Asked to Learn

The experiments eventually changed our question. We first asked which architecture could learn every channel. We later found that an earlier question decides whether learning is possible at all: *does the representation keep the difference that the task asks the model to predict?* If a keypoint encoder turns two different joints into the same nearly empty image, or a semantic-color encoding gives two objects an ambiguous color, the later network no longer has the information needed to tell them apart. Formally, if an encoder maps two different world states x_a and x_b to the same code,

$$E(x_a) = E(x_b), \tag{23}$$

then every later network that sees only that code must treat the two states in the same way, even when their correct answers differ. More layers, more data, a stronger loss, or a larger attention module cannot recover information that was already erased. Our sparse-pixel, semantic-color, shared-VAE, and latent- alignment experiments reached this same limit through different routes.

The useful side of this result is simple. Once every important distinction has a stable address and can be encoded and decoded accurately, it becomes learnable. A model does not need to sense that variable in the same way as a person. A red–green color-blind observer cannot recover color after a sensor has already collapsed red and green into one value. If an auxiliary code records which wavelength class was observed, however, the observer can compare and communicate the two classes without acquiring ordinary color vision. Our heatmap experiments did the same thing for sparse structures. A VAE often erased tiny points on a large empty canvas. Giving each point identity its own heatmap lane turned that identity into a stable address, so the model only had to generate its position and confidence.

This also changed how we understood modality. RGB, depth, pose, material, object structure, audio, and language can describe the same event, but their usual encodings place that information in very different coordinate systems. The difference may lie in value range, spatial layout, density, abstraction, or the meaning assigned to each coordinate. From the viewpoint of a world snapshot, these are different fields of one state. They do not need identical latent coordinates. They need faithful codecs, stable field identities, and a learned relation that says how a known field constrains a requested field. In short:

> *Modality is often distance in representation, not distance in the world.*

This idea can be written compactly. Let a world snapshot contain typed fields $W = \{X_1, \ldots, X_m\}$, and let every field have an encoder and decoder (E_i, D_i). A task chooses the fields already known, $C \subseteq W$, and the fields to be produced, $Y \subseteq W$, then learns

$$p(E_Y(Y) \mid E_C(C),\, s)\,, \tag{24}$$

where s may contain text or other context. In ordinary language, the model generates the requested fields from the known fields. Generation, estimation, control, reconstruction, classification, and structured prediction differ in which fields are known, which are requested, and which direction connects them. Adding a field expands what the model can read and write. Connecting existing fields creates new tasks without requiring a new learning principle.

Large language models provide a useful analogy. A question, a document, a class label, a program, and an answer can all be written as tokens. Once they share that writable form, many language tasks become "given these tokens, generate the next tokens." Computer vision has strong generators, but depth, skeletons, materials, object structures, and sound have usually been treated as outputs of different systems. WSG makes these answers writable as named world fields. The Unified Law supplies the prompt–answer grammar: known world fields form the prompt, and requested world fields form the generated answer.

Vision adds a difficulty that text often hides. Two visual fields may be aligned pixel by pixel and still use incompatible latent coordinates. A codec designed for natural

RGB images may also erase sparse points or distort physical values. Visual unification therefore needs two ingredients. First, every field must cross the input–output boundary without losing the information required by the task. Second, the generator must know how to produce a valid sample in the requested target domain. The codec and sparse experiments revealed the first requirement; the failed external bridges revealed the second.

The resulting scaling strategy has two practical steps. First, enlarge the set of world information that the system can receive and emit faithfully. Second, learn lightweight directed paths among those fields while reusing a shared generative prior. The first step makes the information available; the second makes it usable. This suggests a general rule for foundation-model development:

> *Before scaling what a model knows, scale what the model can faithfully express.*

The same reasoning can extend beyond vision. If a foundation model can receive and emit a new variable accurately, the new variable can become a condition, a target, or both. A tokenizer adds something the model can read or write; a control branch learns how one variable guides another; a router selects the relation required by the current request. Together they give a direct sequence: preserve the information, learn the relation, and then scale the capability.

14.10. What the Experiments Taught Us about Modalities

Visual information is relatively low in abstraction, while textual information is highly abstract; the useful learning form changes with this progression. During development, we gradually realized that moving from visual information to textual information is a process from low abstraction to high abstraction. An image retains concrete location, color, geometry, and texture; text discards most of those sample-specific values and keeps a compact semantic description. The learning objective changes in the same direction, from MSE-like reconstruction of continuous values toward CE-like selection of semantic symbols. A compact, low-channel SD-style VAE latent [30] has already removed much of the spatial detail, so conversion involving it behaves more like prediction. A higher-channel FLUX latent or representation-autoencoder feature retains more visual structure and is consequently better suited to conditional image generation, masking, and spatial translation.

This view places visual fields and text on one representational continuum. If the target exposes more low-level variables, part of the work normally performed by the VAE moves into the generation process; a strong generator can often carry that work. If the same information is converted into reliable semantic tokens, the task moves toward categorical prediction, and greater abstraction can help when sufficient data are available. In this sense, a VAE is the visual counterpart of a tokenizer: it decides which distinctions become the model's symbols. It also divides the work between codec and generator. A compact codec performs more abstraction before learning begins; a less compressed target leaves more low-level structure for the generator to model. Neither choice is universally better. The suitable choice depends on whether the task requires exact continuous reconstruction, spatial translation, or semantic prediction.

Our experiments also found no automatic connection between representations at different abstraction levels. Equal tensor shapes, similar distributions, or paired observations did not give their coordinates the same meaning. Whitening, interpolation, invertible alignment, VQ codes, joint embeddings, flow matching, and diffusion all preserved this mismatch in different forms. We call this the *principle of modal incompatibility*: independently learned modal spaces do not contain a ready-made path between corresponding samples. They do not become mutually readable until a sample-level relation and a usable target decoder connect them. This explains why two encoders can emit arrays of the same shape and scale while a small bridge still cannot translate between them.

A condition constrains only the information it contains, so reversing the direction of a task can change its difficulty completely. Text can precisely narrow the region of a target manifold that should be generated. This helps denoising because the pretrained generator supplies appearance while text selects the intended semantic region. A dense condition such as depth, pose, edges, or material carries strong spatial information, but it does not necessarily specify the appearance variables missing from that condition. A weak semantic condition can be easy to use because it selects within knowledge already stored by the generator. A strong modal condition carries many exact values, yet those values help only when the architecture preserves their spatial identity and gives them a direct path into generation. Strong information is therefore not automatically strong supervision.

The direction of conversion therefore changes the problem. RGB usually constrains depth strongly, so RGB-to-depth is relatively close to prediction. Depth-to-RGB remains one-to-many because depth alone does not determine texture, color, lighting, identity, or background. Without text, noise, style, or other information that selects among these possibilities, velocity-field supervision receives many incompatible valid RGB targets and, in our bridge experiments, often averaged them or failed to reach a useful RGB sample. A condition can guide only the information it contains; direction and missing information must therefore remain part of the task definition.

Dense and sparse fields need different boundaries be-

cause they lose different information when encoded. Our experiments gave a simple working picture. Dense visual data can usually condition another dense field, and a dense observation can be used to estimate a sparse structure. The reverse direction is harder: a sparse structure does not contain all the values needed to produce a dense appearance. Converting the sparse structure into heatmaps or fixed visual lanes makes it easier for an image generator to read, but it does not create texture, illumination, identity, or other information that was never present.

The same distinction explains why one universal image codec performed poorly. Dense fields can be compressed by a VAE, discretized by a VQ codec, or sometimes used directly in pixel space. A sparse field placed on a mostly empty image is dominated by its background, and a tiny reconstruction error can erase a point or change its identity. We therefore tested four practical boundaries: visualize sparse values as heatmaps, predict an ordered sequence, predict the original parameters directly, or use a one-shot structured decoder. Fixed heatmap slots worked because they turned identity into an address and left the generator to predict location and confidence. This did not create an exception to the generative formulation; it showed that different information structures may need different codecs before they enter the same formulation.

Density is not the only distinction. Two fields may cover only partly overlapping facts about the scene. Removing empty regions or constructing a path through the informative regions can reduce wasted computation, but it cannot create values that the condition never observes. A useful practical test is therefore: can the field be converted into an RGB-like representation and decoded back with the distinctions required by the task still intact? If yes, it can often reuse an image generator directly. If no, a sequence, parameter, token, waveform, or other target-native boundary is needed. Our heatmap, object, body-parameter, categorical, and audio experiments are instances of this rule.

The four WSM interfaces are the four directional roles of RGB, text, and additional world information, not arbitrary task labels. Vision, language, and audio are treated as large domains because their usual spaces, abstraction levels, codecs, and pretrained priors are different. Their primary tasks can nevertheless be written in a generative form and can eventually participate in one broader model once each domain can be encoded and decoded faithfully.

The basic capability of an image generator is narrow RGB generation. Generalized visual generation allows the target to be depth, pose, material, motion, topology, or other world information. It does not require one dense joint generator to emit every variable simultaneously: visual chains can compose separately acquired relations and create a requested combination only when it is needed. This mattered in our experiments: joint generation coupled memory and optimization across all outputs, while a chain could reuse independently learned relations. A genuinely different domain such as audio first needs its own faithful codec and target prior, after which it can join the same conditional-generation framework.

Audio illustrates this point. A spectrogram can be handled as a dense visual field, learned audio tokens can be handled as a sequence, and a waveform can be generated by a waveform-native model. These forms carry different inductive biases and should not be mixed into image channels simply because they can be stored in a tensor. Once timing, codec, and target semantics are explicit, however, audio can participate in the same condition–target grammar.

The familiar visual task names then become simple directional cases. Producing additional world information from RGB is estimation. Producing RGB from text or additional world information is generation or conditional control. Reconstruction is a special case of estimation; rerendering is a special case of conditional control in which RGB and another condition are both provided. Spatially aligned world fields can be translated into one another, and using RGB together with property A to recover property B is inverse rendering. These relations directly produce the four WSM interfaces.

Accurate encoding and decoding establish representational access; cross-field use establishes mastery. A codec can read an image and reproduce the same pixels without learning what those pixels imply for depth, motion, material, sound, language, or any other field. This is analogous to memorizing and repeating a statement without understanding its content. An accurate private round trip measures whether a representation can be stored and recovered. Mastery is demonstrated when the backbone uses that representation to infer, generate, or control another one.

What we need from a foundation-model backbone is stronger. A mastered field should be usable as a condition, a target, or an intermediate representation: one field should be able to produce several requested fields, several known fields should be able to determine one requested field, and different fields should participate in composed or arbitrary directed conversions. In this operational sense, mastery is demonstrated by relations, not by identity reconstruction alone. A modality becomes part of the model's usable knowledge only when the model can do something with it beyond returning what it was given.

This distinction also clarifies what a foundation model should seek. It should first acquire basic representations through which vision, language, audio, and other descriptions of the world can be read and written faithfully. It should then learn relations among those representations so that they can serve as inputs and outputs of one-to-one, one-to-many, many-to-one, and composed tasks. Encoding many modalities into tensors is therefore only the represen-

tational boundary of a general model; making them mutually usable is the capability.

Training one model that jointly acquires all such relations at once is often computationally impractical. The Unified Law instead describes incremental expansion from a backbone that already masters part of this representational space: preserve a new field, connect it to existing fields, and install the new directed capability without relearning the whole model. The formal distinction between representational accessibility and operational mastery is summarized in Sec. 7.5.

Which experiments failed, and why. Read against the requirements above, the failures in Table 18 fall into three groups, and each group was repaired only when its missing condition was restored, never by more parameters or longer training.

The representation erased a distinction. A shared VAE erased tiny points on a large empty canvas. Continuous output heads averaged neighboring IDs and confidences, and sparse categorical codecs kept identities better yet still removed points or changed IDs after small decoding errors. A patch/MLP codec removed the empty canvas and with it the spatial organization that a convolution had obtained for free. Semantic-color heatmaps preserved locations while color errors reconnected the wrong joints, and dense pseudocolor object masks separated what a box slot kept together. The same failure occurred on the condition side: deep feature injection into a frozen video generator delivered evidence in a form its blocks could not read, so motion changed while appearance degraded. Heatmaps with fixed person–keypoint slots, typed sequences, object atoms, and named custom fields repaired this group by giving each required distinction a stable address.

Each field lived in its own space. Independent channel VAEs reconstructed individual fields while giving every modality its own latent space. Spaces trained separately in this way are heterogeneous, because their structures differ, and heteronymous, because even two points at corresponding positions of similar-looking structures carry unrelated meanings. Whitening, learnable invertible alignment, and joint-embedding objectives matched shapes and global statistics while individual samples still decoded incorrectly. The shadow-variable route made the failure especially clear: a private RGBD code achieved a nearly perfect round trip yet lost the ability to complete depth from the native RGB latent, and a small normalization change made RAC latents unreadable by an existing PerFlow generator although their own round trip looked correct. External latent bridges—the direct "God Bridge," attention and Transformer bridges, VQ bridges, and conditional latent diffusion—had no shared reference from which to start and could not close the decoded gap. Improving the path did not help either: optimal transport, reflow, intermediate supervision, and noise conditioning made trajectories shorter, straighter, or more diverse, but did not supply the target semantics missing from the bridge. This group was repaired only when all fields were written into one space that the generator already reads and writes: pseudo-RGB serialization through the RGB codec, which may lose some precision but keeps every field in a common space. An encoder that writes every modality into one space and decodes it accurately would serve the same role; the shared representation itself becomes the bridge, even when the codec does not understand what it encodes. Conversion then moved inside a pretrained generator whose prior covered that space, first in video backbones and then in the FLUX-based TunnelingNet routes.

Capabilities shared parameters. Widening a mature RGB codec let every new field alter the shared latent. Joint video lanes interfered once groups reached about four targets, fused conditions were barely trainable, and packed LoRAs raised aggregate throughput by only about two times for eight relations because target-side computation and gradients still grew with each relation. Shared-target LoRAs let every new condition update the same parameters, so adding or repairing one capability could change all previous ones. Separately acquired tunnels repaired this group by storing each condition–target relation in its own small path inside the shared generator.

Across all three groups, apparent success was judged only by complete inference. Low latent losses, partial denoising previews, private round trips, and aligned global statistics repeatedly looked successful while full generation produced black images, fragmented categories, missing structures, or codes unreadable by the native decoder. Section 7.5 states these groups as the preservation, coverage, and installation conditions of the Unified Law and lists the predictions they imply.

14.11. Formal Lessons from the Experiments

The experiments above produced a set of measurable lessons. The following definitions state those lessons in a form that can be tested on new models, representations, and tasks.

K_1. **Information preservation is decided at the model boundary.** Let T map known fields C to requested fields Y, and let $s = (s_C, s_Y)$ be the representation presented to the model. Write $y_a \sim_T y_b$ when two targets are equivalent under the task's native answer and evaluation. A task-sufficient representation must satisfy

$$\begin{aligned} s_C(c_a) = s_C(c_b) &\Rightarrow p_T(Y \mid c_a) = p_T(Y \mid c_b), \\ s_Y(y_a) = s_Y(y_b) &\Rightarrow y_a \sim_T y_b. \end{aligned} \tag{25}$$

The first line says that collapsing two conditions is harmless only when they have the same correct answer distribution. The second says that collapsing two targets is harmless only

when the task itself does not need to distinguish them. Once either implication fails, every downstream model receives the same code for two cases that require different behavior. More parameters, data, or training cannot reconstruct the erased distinction; the remedy is a better boundary. Fixed heatmap lanes, typed sequences, object atoms, and named fields worked because they restored such addresses.

K_2. Private reconstruction does not establish mutual readability. For field X_i with encoder–decoder pair (E_i, D_i), define its private round-trip error and its directed cross-field error by

$$\begin{aligned} r_i &= \mathbb{E}\,\ell_i(D_i(E_i(X_i)), X_i)\,, \\ e_{i\to j}(h) &= \mathbb{E}\,\ell_j(D_j(h(E_i(X_i))), X_j)\,. \end{aligned} \tag{26}$$

Low r_i and r_j, equal tensor shapes, and matched latent statistics do not imply low $e_{i\to j}$ for a small bridge h. Two private codecs may reconstruct perfectly while assigning unrelated meanings to corresponding coordinates. Whitening, invertible alignment, joint embeddings, and the shadow-variable route repeatedly produced this pattern. Mutual readability requires sample-level correspondence, stable field identities, and a target decoder or generator that already gives those coordinates a usable meaning.

K_3. Target abstraction determines how much new output behavior the backbone must learn. For representation Z of underlying field X, define

$$\alpha(Z; X) = H(X \mid Z). \tag{27}$$

Larger α means that more sample-specific information has been discarded before task learning begins. A detailed continuous representation retains location, geometry, color, and texture and therefore favors reconstruction, masking, or spatial translation. A highly abstract representation keeps a smaller semantic choice and behaves more like a categorical prediction target. This is why an image and a sentence describing it occupy different positions on one abstraction continuum, and why MSE-like reconstruction gradually gives way to CE-like prediction. A VAE acts as a visual tokenizer: whatever it removes becomes uncertainty that the generator must supply.

Consider a backbone that is already accustomed to producing dense RGB images. It can continue producing dense outputs from many kinds of conditions. The condition may be text, depth, pose, or another sparse or abstract input; once encoded, it only needs to steer a generation process the backbone already knows. Learning can therefore remain fast even when the condition leaves some image details unresolved. Asking the same backbone to produce text, keypoints, categories, or another sparse and abstract target is a larger change. The model must learn what the new output elements mean, where they are placed, and how they are decoded. The target therefore affects acquisition more strongly than the condition: a familiar target permits many kinds of conditions, while an unfamiliar target requires new output behavior.

This separates two sources of difficulty. A geometry-to-RGB task may leave color, texture, and illumination unresolved yet learn quickly because the backbone already knows how to generate RGB. An RGB-to-structure task may be more fully determined by its condition and still require a new sparse or symbolic output boundary. Directional uncertainty in K_4 measures what the condition leaves unknown; target knowledge in K_6 measures how much of the requested output behavior the backbone already possesses. Both affect learning difficulty.

K_4. Directional difficulty is the information missing from the condition. Define the unresolved information of a directed task by

$$\mu(C \to Y \mid t) = H(Y \mid C, t), \tag{28}$$

where t is optional text or other context. In general, $H(Y \mid C, t) \neq H(C \mid Y, t)$, so reversing two fields creates a different task. RGB often constrains depth, while depth leaves RGB color, texture, illumination, identity, and background unresolved. Text helps only by reducing the missing target information,

$$H(Y \mid C) - H(Y \mid C, t) = I(Y; t \mid C). \tag{29}$$

A condition can therefore guide only the information it contains and the route preserves. High dimensionality is not automatically strong supervision, and a one-to-many reverse direction needs text, noise, another field, or a target prior to supply its missing choices.

K_5. Field density determines whether an image-like boundary is appropriate. Let $\rho(Y)$ be the fraction of spatial positions occupied by a sparse target, and let d_T be its task-native error after decoding. Under an unweighted pixel loss, an empty prediction can obtain

$$\ell_{\mathrm{pix}}(0, Y) = \mathcal{O}(\rho(Y)) \tag{30}$$

while d_T remains maximal because every point or identity is missing. This is why a dense-image VAE can appear numerically successful while erasing sparse structures. Heatmap lanes, fixed slots, sequences, direct parameters, and one-shot structured decoders change which errors receive weight and which identities have stable addresses. They preserve sparse information; they do not invent appearance or physical values absent from the condition.

K_6. A better path cannot replace missing target knowledge. Let $\ell_{\mathrm{path}}(P)$ measure the length, curvature, or transport cost of a path P. When the coverage condition $\kappa(T, G, s) = 0$, the bridge experiments showed

$$\ell_{\mathrm{path}}(P_1) < \ell_{\mathrm{path}}(P_2) \quad \not\Rightarrow \quad Q_T(P_1) > Q_T(P_2) \tag{31}$$

when target-domain knowledge was absent. Optimal transport, reflow, intermediate supervision, noise conditioning, and repeated correction could make a path shorter, straighter, or more robust without teaching valid target texture or structure. Path design controls how generation moves; the target prior determines what valid destinations look like. Once that prior exists, a lightweight route can redirect it efficiently.

K_7**. Independent capability installation is measured by interference.** For an installed task i and newly added route j, define

$$\Delta Q_{i \leftarrow j} = Q_i(G, \Phi) - Q_i(G, \Phi \cup \{\phi_j\}). \tag{32}$$

Independent installation requires $\Delta Q_{i \leftarrow j} = 0$ up to evaluation tolerance, even though every route remains part of one shared model and runtime. Joint lanes, fused conditions, packed LoRAs, and shared-target LoRAs coupled gradients, memory, schedules, or later revisions. Separately optimized tunnels restrict each update to ϕ_j. For N installed routes,

$$P_{\text{resident}}(N) = P_G + \sum_{i=1}^{N} |\phi_i|, \qquad P_{\text{active}}(j) = P_G + |\phi_j|. \tag{33}$$

The resident capability collection grows with the number of installed relations, while task-specific active parameters and marginal acquisition cost remain bounded by one selected route. This distinction, not a claim of constant total storage, explains why thousand-task expansion becomes practical.

K_8**. Success is a property of the decoded terminal object.** Let $\mathcal{L}_{\text{sur}}$ denote a training loss, latent alignment, or intermediate preview, and let Q_T evaluate the complete decoded output in the task's native space. For two candidate designs a and b, the experiments repeatedly gave

$$\mathcal{L}_{\text{sur}}(a) < \mathcal{L}_{\text{sur}}(b) \quad \not\Rightarrow \quad Q_T(a) > Q_T(b). \tag{34}$$

Low surrogate loss can accompany black images, missing points, fragmented categories, or codes unreadable by the native decoder. Intermediate quantities diagnose a route; they do not establish output quality. The complete trajectory, decoder, and target-native evaluator determine Q_T.

The remaining families supplied six more specific diagnostics. They refine the eight core principles without changing the law's statement.

K_9**. Local path accuracy and rollout accuracy are different.** For an n-step update with local errors ε_k and state-transition Lipschitz factors λ_k, terminal deviation can accumulate as

$$\varepsilon_{\text{roll}}^{(n)} \lesssim \sum_{k=1}^{n} \left(\prod_{q=k+1}^{n} \lambda_q \right) \varepsilon_k. \tag{35}$$

Random-time training, one-step endpoint losses, and intermediate supervision change different terms in this expression. A small local velocity error does not ensure a small full-trajectory error; matched integration and full decoded evaluation are required.

K_{10}**. A discrete code needs stable coverage and a meaningful metric.** For codebook size K, let $u = |\mathcal{K}_{\text{used}}|/K$ be utilization and let δ_{code} measure how much code meanings move during training. High utilization with large δ_{code} gives a moving classification target, while frozen codes can give stable labels with poor coverage. Moreover, nominal IDs have no default ordinal geometry:

$$d_{\text{semantic}}(k_a, k_b) \not\propto |k_a - k_b|. \tag{36}$$

This separates codebook health, label stability, and cross-field predictability.

K_{11}**. Diversity and fidelity are independent.** Generated diversity is measured directly by $H(\hat{Y} \mid C, t)$. Noise identities, CVAE variables, best-of-k selection, target noising, and guidance can increase or expose it, but

$$H(\hat{Y} \mid C, t) \uparrow \quad \not\Rightarrow \quad Q_T \uparrow. \tag{37}$$

Stochasticity represents unresolved choices; it does not teach the model which samples are valid in the target domain.

K_{12}**. Engineering speedups and semantic progress must be accounted for separately.** Write the measured acquisition cost as

$$\mathcal{C}_{\text{total}} = \mathcal{C}_{\text{data}} + \mathcal{C}_{\text{codec}} + \mathcal{C}_{\text{route}} + \mathcal{C}_{\text{target}}. \tag{38}$$

Caching reduces the first two terms; packing can share part of $\mathcal{C}_{\text{route}}$; neither operation necessarily reduces translation error or the target-specific gradients in $\mathcal{C}_{\text{target}}$. This distinction prevented throughput improvements from being mistaken for a solved mapping.

K_{13}**. The best routing and objective depend on the declared task.** For permitted route family $\mathcal{M}$ and budget B, define

$$\phi_T^* = \arg \max_{\phi \in \mathcal{M},\, \mathcal{C}(\phi) \le B} Q_T(G, \phi). \tag{39}$$

The experiments give $\phi_T^* \neq \phi_{T'}^*$ across interfaces: structured targets favored separately normalized evidence routing, RGB retained an advantage from native joint attention, and categorical outputs benefited from explicit symbolic recovery. A unified task grammar therefore does not require one identical internal route or loss for every information structure.

K_{14}**. Installed relations form executable compositions.** For installed directed kernels $p_{\phi_1}, \ldots, p_{\phi_L}$, a visual chain is the explicit composition

$$p(Y_L \mid C_0) = \int \prod_{\ell=1}^{L} p_{\phi_\ell}(Y_\ell \mid Y_{\ell-1})\, dY_{1:L-1}. \tag{40}$$

Separate route provenance is retained at every edge, so a chain can be assembled without joint retraining. Its quality remains an empirical property of the component routes and their intermediate representations, not an automatic consequence of combinatorial path count.

These fourteen parameterized principles retain the plain conclusions of the development record while making their terms observable. Together they give the sequence that ultimately produced the Unified Law: preserve what the task must distinguish; place fields in mutually usable representations; measure the missing information in the declared direction; choose a boundary suited to the field's abstraction and density; reuse a prior that knows valid targets; install each relation without disturbing previous ones; and judge success only after complete inference. The last six additionally formalize rollout error, discrete codes, diversity, resource accounting, task-dependent routing, and runtime composition. Section 7.5 extracts the smallest set needed to state and refute the law.

The following traceability table closes the loop between the experimental record and the formal statements. Each of the 60 rows covers the named variants counted inside that family; the 52 most consequential variants are also unpacked individually in Table 19. Thus the 198 currently indexed named variants enter the formal account without pretending that every engineering variant requires a new symbol or a separate law.

Table 20. **Formal traceability of all 60 experimental families.** The right column identifies the parameterized conclusions supported or qualified by every family. Table 21 then opens the family counts into 198 individual variants.

Experimental family	Formal knowledge
Expanded RGB VAE	$\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_7$
Frozen RGB path	$\mathsf{K}_1, \mathsf{K}_7$
Progressive TinyVAE expansion	$\mathsf{K}_1, \mathsf{K}_3, \mathsf{K}_5, \mathsf{K}_{12}$
Offset-routing schemes D–F	$\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_6, \mathsf{K}_{13}$
LayerDiffuse-style auxiliary path	$\mathsf{K}_1, \mathsf{K}_7, \mathsf{K}_{13}$
Independent channel VAE	$\mathsf{K}_1, \mathsf{K}_2$
Sparse categorical VAE	$\mathsf{K}_1, \mathsf{K}_5, \mathsf{K}_{10}$
Sparse patch/MLP VAE	$\mathsf{K}_1, \mathsf{K}_5, \mathsf{K}_{13}$
Cached paired latents	K_{12}
Latent objectives and masking	$\mathsf{K}_8, \mathsf{K}_{13}$
Seven local interpolation rules	$\mathsf{K}_2, \mathsf{K}_6$
Direct “God Bridge”	$\mathsf{K}_2, \mathsf{K}_6$
Attention bridge	$\mathsf{K}_2, \mathsf{K}_6, \mathsf{K}_{13}$
Bidirectional CNN bridge	$\mathsf{K}_2, \mathsf{K}_4$
Whitening and fixed normalization	$\mathsf{K}_2, \mathsf{K}_6$
Learnable invertible alignment	$\mathsf{K}_2, \mathsf{K}_6$
Joint embedding and JEPA-style alignment	$\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_3$
Multi-step RF repair	$\mathsf{K}_6, \mathsf{K}_9, \mathsf{K}_{12}$
Random-time RF	$\mathsf{K}_8, \mathsf{K}_9$
Endpoint and intermediate losses	$\mathsf{K}_8, \mathsf{K}_9, \mathsf{K}_{13}$
Minibatch optimal transport	$\mathsf{K}_6, \mathsf{K}_9$
Reflow and teacher-path correction	$\mathsf{K}_6, \mathsf{K}_9$
Noise ranking and noise conditioning	$\mathsf{K}_4, \mathsf{K}_{11}$
VQ bridge with CE	$\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_{10}$
VQ utilization variants	$\mathsf{K}_2, \mathsf{K}_{10}$
MLP index-MSE bridge	$\mathsf{K}_1, \mathsf{K}_{10}, \mathsf{K}_{13}$
Transformer bridge	$\mathsf{K}_2, \mathsf{K}_6, \mathsf{K}_{13}$
Pure velocity field, CFG, and condition dropout	$\mathsf{K}_4, \mathsf{K}_6, \mathsf{K}_{11}$
CVAE-style prior–posterior bridge	$\mathsf{K}_4, \mathsf{K}_6, \mathsf{K}_{11}$
Best-of-k endpoint sampling	$\mathsf{K}_6, \mathsf{K}_8, \mathsf{K}_{11}$
Target-noised training	$\mathsf{K}_6, \mathsf{K}_9, \mathsf{K}_{11}$
Style-latent conditioning	$\mathsf{K}_4, \mathsf{K}_6, \mathsf{K}_{11}$
Reverse-noise encoding	$\mathsf{K}_6, \mathsf{K}_9$
Conditional latent diffusion	$\mathsf{K}_6, \mathsf{K}_9, \mathsf{K}_{13}$
Rectified auto-coder paths	$\mathsf{K}_2, \mathsf{K}_6, \mathsf{K}_9$
Shadow-variable and perturbed-latent routes	$\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_6$
Video-backbone branches	$\mathsf{K}_6, \mathsf{K}_7, \mathsf{K}_{13}$
Joint video-lane strategies	$\mathsf{K}_7, \mathsf{K}_{12}, \mathsf{K}_{13}$
FLUX/Omini evidence placement	$\mathsf{K}_4, \mathsf{K}_{13}$
Task-agnostic flow objectives	$\mathsf{K}_8, \mathsf{K}_9, \mathsf{K}_{13}$
Dense flow objective levers	$\mathsf{K}_8, \mathsf{K}_{13}$
Packed multi-target LoRA acquisition	$\mathsf{K}_7, \mathsf{K}_{12}$
Shared-target multi-condition LoRA	$\mathsf{K}_4, \mathsf{K}_7, \mathsf{K}_{12}$
NJA/MTA TunnelingNet routes	$\mathsf{K}_6, \mathsf{K}_7, \mathsf{K}_{13}$
Semantic heatmap and inverse codec	$\mathsf{K}_1, \mathsf{K}_5$
Person–keypoint slot heatmaps	$\mathsf{K}_1, \mathsf{K}_5$
Body-part and full-body heatmaps	$\mathsf{K}_1, \mathsf{K}_5, \mathsf{K}_{12}$
Direct sparse sequence heads	$\mathsf{K}_1, \mathsf{K}_5, \mathsf{K}_{13}$
ScoreHMR WSV structure route	$\mathsf{K}_6, \mathsf{K}_{13}$
Atomic object-box slots	$\mathsf{K}_1, \mathsf{K}_5, \mathsf{K}_{13}$
Object BFS tree	$\mathsf{K}_1, \mathsf{K}_5$
Point-cloud skeleton initialization	$\mathsf{K}_5, \mathsf{K}_6, \mathsf{K}_{13}$
Sparse and mixed audio routes	$\mathsf{K}_1, \mathsf{K}_3, \mathsf{K}_4, \mathsf{K}_{13}$
Dense categorical-mask pilots	$\mathsf{K}_5, \mathsf{K}_8, \mathsf{K}_{10}$

Continued on the next page.

Table 20 continued.

Experimental family	Formal knowledge
Categorical objective modes	$\mathsf{K}_8, \mathsf{K}_{10}, \mathsf{K}_{13}$
Categorical region refiners	$\mathsf{K}_8, \mathsf{K}_{10}, \mathsf{K}_{13}$
Universal Pose custom Head	$\mathsf{K}_1, \mathsf{K}_5, \mathsf{K}_{13}$
Pose-detail objective pilots	$\mathsf{K}_5, \mathsf{K}_8, \mathsf{K}_{13}$
Named custom-variable routes	$\mathsf{K}_1, \mathsf{K}_7, \mathsf{K}_{13}$
Runtime route composition	$\mathsf{K}_7, \mathsf{K}_{14}$

Table 21. **Variant-level evidence and formal traceability.** The 198 currently indexed methodological variants are listed independently. Each row records the intervention actually distinguished in the development record, the direct observation retained from it, and the formal principle(s) it supports. Engineering-only repairs, repeated seeds, and ordinary hyperparameter sweeps remain excluded.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
1	Expanded RGB VAE	Direct expanded RGB input/output boundary	Directly widening the mature RGB boundary disturbed its latent; added information and preserved RGB competed.	K_1, K_2, K_7
2	Frozen RGB path	Frozen native RGB path with trainable outer layers	Freezing preserved RGB but left the added field too little influence.	K_1, K_7
3	Progressive TinyVAE expansion	Hidden-width expansion in 16-channel stages	Extra codec capacity helped dense reconstruction but did not solve sparse structure or full-tensor scaling.	K_1, K_3, K_5, K_12
4	Progressive TinyVAE expansion	Five-channel latent stage	Extra codec capacity helped dense reconstruction but did not solve sparse structure or full-tensor scaling.	K_1, K_3, K_5, K_12
5	Progressive TinyVAE expansion	Six-channel latent stage	Extra codec capacity helped dense reconstruction but did not solve sparse structure or full-tensor scaling.	K_1, K_3, K_5, K_12
6	Progressive TinyVAE expansion	Seven-channel latent stage	Extra codec capacity helped dense reconstruction but did not solve sparse structure or full-tensor scaling.	K_1, K_3, K_5, K_12
7	Progressive TinyVAE expansion	Eight-channel latent stage	Extra codec capacity helped dense reconstruction but did not solve sparse structure or full-tensor scaling.	K_1, K_3, K_5, K_12
8	Progressive TinyVAE expansion	Full 256-channel WSG training	Extra codec capacity helped dense reconstruction but did not solve sparse structure or full-tensor scaling.	K_1, K_3, K_5, K_12
9	Offset-routing schemes D--F	Single-depth direct offset injection	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
10	Offset-routing schemes D--F	Multi-depth direct offset injection	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
11	Offset-routing schemes D--F	Decoder-entrance injection	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
12	Offset-routing schemes D--F	Late-decoder injection	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
13	Offset-routing schemes D--F	Four-scale encoder--decoder jump links	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
14	Offset-routing schemes D--F	Decoder-predicted offset recovery	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
15	Offset-routing schemes D--F	Fixed-ratio offset recovery	Injection position changed the information available to the decoder; better round trips did not ensure a self-contained latent.	K_1, K_2, K_6, K_13
16	LayerDiffuse-style auxiliary path	Auxiliary encoder lane	A protected auxiliary path carried limited extra information but lacked bandwidth for a growing world representation.	K_1, K_7, K_13
17	LayerDiffuse-style auxiliary path	Auxiliary decoder lane	A protected auxiliary path carried limited extra information but lacked bandwidth for a growing world representation.	K_1, K_7, K_13
18	LayerDiffuse-style auxiliary path	Transparency-style protected auxiliary path	A protected auxiliary path carried limited extra information but lacked bandwidth for a growing world representation.	K_1, K_7, K_13
19	Independent channel VAE	Per-channel continuous VAE	Private codecs reconstructed fields but created mutually unreadable coordinate systems.	K_1, K_2
20	Independent channel VAE	Field-specific normalized VAE	Private codecs reconstructed fields but created mutually unreadable coordinate systems.	K_1, K_2
21	Independent channel VAE	Independent field latent space	Private codecs reconstructed fields but created mutually unreadable coordinate systems.	K_1, K_2
22	Sparse categorical VAE	Categorical sparse decoder	Discrete supervision preserved symbols better, while small spatial or class errors still destroyed sparse identities.	K_1, K_5, K_10
23	Sparse categorical VAE	Integer-bin output representation	Discrete supervision preserved symbols better, while small spatial or class errors still destroyed sparse identities.	K_1, K_5, K_10
24	Sparse categorical VAE	Cross-entropy sparse objective	Discrete supervision preserved symbols better, while small spatial or class errors still destroyed sparse identities.	K_1, K_5, K_10
25	Sparse categorical VAE	Hybrid CE--MSE sparse objective	Discrete supervision preserved symbols better, while small spatial or class errors still destroyed sparse identities.	K_1, K_5, K_10
26	Sparse patch/MLP VAE	Patch/MLP sparse codec	Removing empty pixels reduced imbalance but removed spatial organization; coordinates restored only part of it.	K_1, K_5, K_13
27	Sparse patch/MLP VAE	Valid-value tokenization	Removing empty pixels reduced imbalance but removed spatial organization; coordinates restored only part of it.	K_1, K_5, K_13
28	Sparse patch/MLP VAE	Coordinate-augmented valid-value tokens	Removing empty pixels reduced imbalance but removed spatial organization; coordinates restored only part of it.	K_1, K_5, K_13

Continued on the next page.

Table 21 continued.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
29	Cached paired latents	Cached paired source/target latents	Caching separated image I/O and codec cost from the scientific cross-space mapping problem.	$K_1 2$
30	Latent objectives and masking	Latent MSE	Numerically better latent optimization did not guarantee correct decoded endpoints.	$K_8, K_1 3$
31	Latent objectives and masking	Latent MAE	Numerically better latent optimization did not guarantee correct decoded endpoints.	$K_8, K_1 3$
32	Latent objectives and masking	Decoded pixel supervision	Numerically better latent optimization did not guarantee correct decoded endpoints.	$K_8, K_1 3$
33	Latent objectives and masking	Random 4-by-4 latent masking	Numerically better latent optimization did not guarantee correct decoded endpoints.	$K_8, K_1 3$
34	Seven local interpolation rules	Nearest-neighbor interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
35	Seven local interpolation rules	RBF interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
36	Seven local interpolation rules	Thin-plate interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
37	Seven local interpolation rules	Barycentric interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
38	Seven local interpolation rules	Local-linear interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
39	Seven local interpolation rules	Manifold interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
40	Seven local interpolation rules	Neural interpolation	No tested local rule made neighborhoods in independently trained latent spaces semantically corresponding.	K_2, K_6
41	Direct God Bridge	Direct bidirectional God Bridge	A compact one-pass mapper could not discover the complete cross-space relation.	K_2, K_6
42	Attention bridge	Self-attention bridge	Global interaction and added width did not repair incompatible endpoint meanings.	$K_2, K_6, K_1 3$
43	Attention bridge	Cross-attention bridge	Global interaction and added width did not repair incompatible endpoint meanings.	$K_2, K_6, K_1 3$
44	Attention bridge	Compact 4M attention bridge	Global interaction and added width did not repair incompatible endpoint meanings.	$K_2, K_6, K_1 3$
45	Bidirectional CNN bridge	RGB-to-field residual CNN	Opposite directions learned at different rates and qualities, making direction part of the task.	K_2, K_4
46	Bidirectional CNN bridge	Field-to-RGB residual CNN	Opposite directions learned at different rates and qualities, making direction part of the task.	K_2, K_4
47	Whitening and fixed normalization	Fixed mean/variance whitening	Matching scale improved conditioning without aligning sample-level semantics.	K_2, K_6
48	Learnable invertible alignment	Learned 1-by-1 invertible projection	Population statistics could align while corresponding samples still decoded incorrectly.	K_2, K_6
49	Learnable invertible alignment	Moment-matching alignment	Population statistics could align while corresponding samples still decoded incorrectly.	K_2, K_6
50	Learnable invertible alignment	Covariance-matching alignment	Population statistics could align while corresponding samples still decoded incorrectly.	K_2, K_6
51	Joint embedding and JEPA-style alignment	Joint-embedding alignment	Shared semantics emerged, but exact modality-specific information still required separate capacity.	K_1, K_2, K_3
52	Joint embedding and JEPA-style alignment	JEPA-style predictive alignment	Shared semantics emerged, but exact modality-specific information still required separate capacity.	K_1, K_2, K_3
53	Multi-step RF repair	Fully unrolled multi-step RF repair	Repeated repair increased memory and instability without supplying missing target knowledge.	$K_6, K_9, K_1 2$
54	Random-time RF	Random-time single-point RF training	Random-time supervision saved memory, while local velocity errors accumulated during full rollout.	K_8, K_9
55	Endpoint and intermediate losses	One-step endpoint loss	Endpoint-aware losses exposed hidden failures, but matched full paths restored substantial cost and could conflict with local fitting.	$K_8, K_9, K_1 3$
56	Endpoint and intermediate losses	Matched multi-step endpoint loss	Endpoint-aware losses exposed hidden failures, but matched full paths restored substantial cost and could conflict with local fitting.	$K_8, K_9, K_1 3$
57	Endpoint and intermediate losses	Checkpointed endpoint integration	Endpoint-aware losses exposed hidden failures, but matched full paths restored substantial cost and could conflict with local fitting.	$K_8, K_9, K_1 3$

Continued on the next page.

Table 21 continued.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
58	Endpoint and intermediate losses	Intermediate-state supervision	Endpoint-aware losses exposed hidden failures, but matched full paths restored substantial cost and could conflict with local fitting.	K_8, K_9, K_{13}
59	Endpoint and intermediate losses	Aligned/outside-space endpoint supervision	Endpoint-aware losses exposed hidden failures, but matched full paths restored substantial cost and could conflict with local fitting.	K_8, K_9, K_{13}
60	Minibatch optimal transport	Minibatch OT pairing	Optimal pairing shortened and uncrossed paths without supplying target semantics.	K_6, K_9
61	Reflow and teacher-path correction	Reflow retraining	Path refinement helped only when the teacher already reached useful endpoints.	K_6, K_9
62	Reflow and teacher-path correction	Teacher-endpoint path correction	Path refinement helped only when the teacher already reached useful endpoints.	K_6, K_9
63	Reflow and teacher-path correction	Repeated correction blocks	Path refinement helped only when the teacher already reached useful endpoints.	K_6, K_9
64	Noise ranking and noise conditioning	Persistent noise ranking	Identifiable random starts increased path diversity but did not create target realism.	K_4, K_{11}
65	Noise ranking and noise conditioning	Explicit noise conditioning	Identifiable random starts increased path diversity but did not create target realism.	K_4, K_{11}
66	VQ bridge with CE	Frozen-codebook VQ bridge with CE	Legal code predictions still suffered from low utilization, moving labels, and cross-space semantic mismatch.	K_1, K_2, K_{10}
67	VQ bridge with CE	Updated-codebook VQ bridge with CE	Legal code predictions still suffered from low utilization, moving labels, and cross-space semantic mismatch.	K_1, K_2, K_{10}
68	VQ utilization variants	Smaller codebook	Healthier code usage did not by itself improve cross-field translation.	K_2, K_{10}
69	VQ utilization variants	Codebook warm-up updates	Healthier code usage did not by itself improve cross-field translation.	K_2, K_{10}
70	VQ utilization variants	Utilization-aware monitoring	Healthier code usage did not by itself improve cross-field translation.	K_2, K_{10}
71	VQ utilization variants	Alternative CE--MSE weighting	Healthier code usage did not by itself improve cross-field translation.	K_2, K_{10}
72	MLP index-MSE bridge	MLP index-MSE prediction	Numerical distance between nominal code IDs was not a semantic metric.	K_1, K_{10}, K_{13}
73	Transformer bridge	Transformer latent bridge	A global sequence model also failed to close the decoded distribution gap.	K_2, K_6, K_{13}
74	Pure velocity field, CFG, and condition dropout	Pure conditional velocity objective	A cleaner conditional path helped toward RGB, while information-poor reverse directions remained difficult.	K_4, K_6, K_{11}
75	Pure velocity field, CFG, and condition dropout	Classifier-free guidance	A cleaner conditional path helped toward RGB, while information-poor reverse directions remained difficult.	K_4, K_6, K_{11}
76	Pure velocity field, CFG, and condition dropout	Condition dropout	A cleaner conditional path helped toward RGB, while information-poor reverse directions remained difficult.	K_4, K_6, K_{11}
77	CVAE-style prior--posterior bridge	CVAE prior--posterior bridge	A stochastic latent represented uncertainty but introduced another difficult prior and target decoder.	K_4, K_6, K_{11}
78	Best-of-k endpoint sampling	Best-of-k endpoint selection	Search tolerated several valid answers but could not repair the generator producing them.	K_6, K_8, K_{11}
79	Target-noised training	Target-noised conditional denoising	Making the target distribution explicit improved diversity but left decoded detail unstable.	K_6, K_9, K_{11}
80	Target-noised training	Target-noised conditional velocity	Making the target distribution explicit improved diversity but left decoded detail unstable.	K_6, K_9, K_{11}
81	Style-latent conditioning	Style-latent conditioning	The style variable exposed missing appearance choices but recreated the bridge/prior problem.	K_4, K_6, K_{11}
82	Reverse-noise encoding	Reverse-noise encoder	A structured inverse start did not make forward target generation reliable.	K_6, K_9
83	Conditional latent diffusion	Conditional DDPM bridge	Changing the noise schedule from RF to diffusion did not remove the missing-target-prior failure.	K_6, K_9, K_{13}
84	Conditional latent diffusion	Conditional DDIM bridge	Changing the noise schedule from RF to diffusion did not remove the missing-target-prior failure.	K_6, K_9, K_{13}
85	Rectified auto-coder paths	One-step RAC	Path regularity, perturbation recovery, and compatibility with a pretrained latent contract were separate properties.	K_2, K_6, K_9

Continued on the next page.

Table 21 continued.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
86	Rectified auto-coder paths	Four-step RAC	Path regularity, perturbation recovery, and compatibility with a pretrained latent contract were separate properties.	K_2, K_6, K_9
87	Rectified auto-coder paths	Fixed-time-grid RAC	Path regularity, perturbation recovery, and compatibility with a pretrained latent contract were separate properties.	K_2, K_6, K_9
88	Rectified auto-coder paths	Random-time-grid RAC	Path regularity, perturbation recovery, and compatibility with a pretrained latent contract were separate properties.	K_2, K_6, K_9
89	Rectified auto-coder paths	Intermediate-state RAC supervision	Path regularity, perturbation recovery, and compatibility with a pretrained latent contract were separate properties.	K_2, K_6, K_9
90	Rectified auto-coder paths	State-noise correction RAC	Path regularity, perturbation recovery, and compatibility with a pretrained latent contract were separate properties.	K_2, K_6, K_9
91	Shadow-variable and perturbed-latent routes	Direct RGBD shadow code	A nearby or privately reconstructable code was not necessarily interchangeable with the native RGB latent.	K_1, K_2, K_6
92	Shadow-variable and perturbed-latent routes	Depth visualized through the official RGB codec	A nearby or privately reconstructable code was not necessarily interchangeable with the native RGB latent.	K_1, K_2, K_6
93	Shadow-variable and perturbed-latent routes	Position-matched bidirectional shadow encoding	A nearby or privately reconstructable code was not necessarily interchangeable with the native RGB latent.	K_1, K_2, K_6
94	Shadow-variable and perturbed-latent routes	Additive perturbation around native RGB latent	A nearby or privately reconstructable code was not necessarily interchangeable with the native RGB latent.	K_1, K_2, K_6
95	Video-backbone branches	LTX deep injection	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
96	Video-backbone branches	Native LTX LoRA	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
97	Video-backbone branches	Modality-tagged Wan LoRA	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
98	Video-backbone branches	UniVidX baseline	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
99	Video-backbone branches	WSG UniVid adapter	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
100	Video-backbone branches	Official-style video control	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
101	Video-backbone branches	Wan 1.3B route	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
102	Video-backbone branches	Fun InP reference route	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
103	Video-backbone branches	VACE reference route	A pretrained video generator supplied a target prior, but only appropriately placed interventions redirected it reliably.	K_6, K_7, K_13
104	Joint video-lane strategies	Native multi-lane generation	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
105	Joint video-lane strategies	Sample-one lane training	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
106	Joint video-lane strategies	Split condition groups	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
107	Joint video-lane strategies	Fused conditions	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
108	Joint video-lane strategies	Merged-lane big tokens	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
109	Joint video-lane strategies	Multimodal lane tags	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
110	Joint video-lane strategies	Multilayer residual control	Fixed small lane sets were feasible, while larger collections became memory-heavy and mutually coupled.	K_7, K_12, K_13
111	FLUX/Omini evidence placement	Target-first native reference	Evidence placement and attention topology changed convergence even with identical data and targets.	K_4, K_13
112	FLUX/Omini evidence placement	Trainable condition-lane reference	Evidence placement and attention topology changed convergence even with identical data and targets.	K_4, K_13
113	FLUX/Omini evidence placement	Dual condition attention	Evidence placement and attention topology changed convergence even with identical data and targets.	K_4, K_13
114	FLUX/Omini evidence placement	Task-type attention switch	Evidence placement and attention topology changed convergence even with identical data and targets.	K_4, K_13

Continued on the next page.

Table 21 continued.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
115	FLUX/Omini evidence placement	Frozen condition-side LoRA	Evidence placement and attention topology changed convergence even with identical data and targets.	K_4, K_13
116	Task-agnostic flow objectives	Plain rectified flow	The high-noise anchor best preserved global structure under fixed validation, while objectives remained task-independent tools.	K_8, K_9, K_13
117	Task-agnostic flow objectives	Direct clean-latent prediction	The high-noise anchor best preserved global structure under fixed validation, while objectives remained task-independent tools.	K_8, K_9, K_13
118	Task-agnostic flow objectives	High-noise anchor	The high-noise anchor best preserved global structure under fixed validation, while objectives remained task-independent tools.	K_8, K_9, K_13
119	Task-agnostic flow objectives	Multi-resolution noise	The high-noise anchor best preserved global structure under fixed validation, while objectives remained task-independent tools.	K_8, K_9, K_13
120	Dense flow objective levers	Low-noise timestep shift	Individually useful loss changes interfered when combined and could encourage mean outputs.	K_8, K_13
121	Dense flow objective levers	Min-SNR weighting	Individually useful loss changes interfered when combined and could encourage mean outputs.	K_8, K_13
122	Dense flow objective levers	Huber loss	Individually useful loss changes interfered when combined and could encourage mean outputs.	K_8, K_13
123	Packed multi-target LoRA acquisition	Full eight-target packed graph	Packing shared some execution but target-specific activations, gradients, and cost still scaled with relation count.	K_7, K_12
124	Packed multi-target LoRA acquisition	Four-target retained-graph chunks	Packing shared some execution but target-specific activations, gradients, and cost still scaled with relation count.	K_7, K_12
125	Packed multi-target LoRA acquisition	Four-target immediate-backward chunks	Packing shared some execution but target-specific activations, gradients, and cost still scaled with relation count.	K_7, K_12
126	Packed multi-target LoRA acquisition	Two-target single-GPU chunks	Packing shared some execution but target-specific activations, gradients, and cost still scaled with relation count.	K_7, K_12
127	Packed multi-target LoRA acquisition	Frozen condition-side packed encoder	Packing shared some execution but target-specific activations, gradients, and cost still scaled with relation count.	K_7, K_12
128	Shared-target multi-condition LoRA	Eight-condition alternating Edge LoRA	A common target did not make distinct condition-to-target relations independently maintainable.	K_4, K_7, K_12
129	Shared-target multi-condition LoRA	Four-condition dense groups	A common target did not make distinct condition-to-target relations independently maintainable.	K_4, K_7, K_12
130	Shared-target multi-condition LoRA	Fused-condition shared route	A common target did not make distinct condition-to-target relations independently maintainable.	K_4, K_7, K_12
131	Shared-target multi-condition LoRA	Thirty-eight-condition expansion	A common target did not make distinct condition-to-target relations independently maintainable.	K_4, K_7, K_12
132	NJA/MTA TunnelingNet routes	Native joint attention route	Small internal residual paths could redirect a shared prior; the preferred attention topology depended on the interface.	K_6, K_7, K_13
133	NJA/MTA TunnelingNet routes	Modality tunneling attention route	Small internal residual paths could redirect a shared prior; the preferred attention topology depended on the interface.	K_6, K_7, K_13
134	NJA/MTA TunnelingNet routes	Zero-initialized residual control	Small internal residual paths could redirect a shared prior; the preferred attention topology depended on the interface.	K_6, K_7, K_13
135	NJA/MTA TunnelingNet routes	Multilayer residual tunnel	Small internal residual paths could redirect a shared prior; the preferred attention topology depended on the interface.	K_6, K_7, K_13
136	Semantic heatmap and inverse codec	Gaussian semantic heatmap	Heatmaps preserved locations better, while semantic color still entangled identity.	K_1, K_5
137	Semantic heatmap and inverse codec	Peak/NMS inverse heatmap codec	Heatmaps preserved locations better, while semantic color still entangled identity.	K_1, K_5
138	Person--keypoint slot heatmaps	Fixed person--keypoint slots	Fixed slots made identity structural and left only location/confidence to generation.	K_1, K_5
139	Person--keypoint slot heatmaps	Three-lane packed slot heatmaps	Fixed slots made identity structural and left only location/confidence to generation.	K_1, K_5
140	Body-part and full-body heatmaps	Body-part partitioned heatmaps	Partitioning reduced identity collisions at the cost of additional lanes and caching.	K_1, K_5, K_12
141	Body-part and full-body heatmaps	Body-only heatmaps	Partitioning reduced identity collisions at the cost of additional lanes and caching.	K_1, K_5, K_12
142	Body-part and full-body heatmaps	Full-body heatmap codebook	Partitioning reduced identity collisions at the cost of additional lanes and caching.	K_1, K_5, K_12
143	Direct sparse sequence heads	Keypoint sequence head	Typed sequences were compact but required explicit ordering, masks, matching, decoding, and native losses.	K_1, K_5, K_13

Continued on the next page.

Table 21 continued.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
144	Direct sparse sequence heads	Pose sequence head	Typed sequences were compact but required explicit ordering, masks, matching, decoding, and native losses.	K_1, K_5, K_13
145	Direct sparse sequence heads	Box sequence head	Typed sequences were compact but required explicit ordering, masks, matching, decoding, and native losses.	K_1, K_5, K_13
146	Direct sparse sequence heads	SMPL parameter head	Typed sequences were compact but required explicit ordering, masks, matching, decoding, and native losses.	K_1, K_5, K_13
147	Direct sparse sequence heads	Audio-token sequence head	Typed sequences were compact but required explicit ordering, masks, matching, decoding, and native losses.	K_1, K_5, K_13
148	ScoreHMR WSV structure route	Full-frame cached ScoreHMR features	A pretrained structural prior learned body parameters quickly from aligned image evidence.	K_6, K_13
149	ScoreHMR WSV structure route	Crop-aligned cached ScoreHMR features	A pretrained structural prior learned body parameters quickly from aligned image evidence.	K_6, K_13
150	Atomic object-box slots	Structured center/size box slots	Atomic center/size slots preserved object membership better than dense pseudocolor masks.	K_1, K_5, K_13
151	Atomic object-box slots	MTA pseudocolor box masks	Atomic center/size slots preserved object membership better than dense pseudocolor masks.	K_1, K_5, K_13
152	Atomic object-box slots	NJA pseudocolor box masks	Atomic center/size slots preserved object membership better than dense pseudocolor masks.	K_1, K_5, K_13
153	Object BFS tree	Parent-relative BFS serialization	Explicit parent order and traversal represented variable object topology better than fixed human joint IDs.	K_1, K_5
154	Object BFS tree	Absolute-coordinate BFS serialization	Explicit parent order and traversal represented variable object topology better than fixed human joint IDs.	K_1, K_5
155	Point-cloud skeleton initialization	K-means point-cloud anchors	Analytical geometry supplied a useful starting skeleton; unconstrained learned refinement remained unstable.	K_5, K_6, K_13
156	Point-cloud skeleton initialization	MST skeleton initialization	Analytical geometry supplied a useful starting skeleton; unconstrained learned refinement remained unstable.	K_5, K_6, K_13
157	Point-cloud skeleton initialization	Diffusion skeleton refinement	Analytical geometry supplied a useful starting skeleton; unconstrained learned refinement remained unstable.	K_5, K_6, K_13
158	Point-cloud skeleton initialization	Anchor-correction refinement	Analytical geometry supplied a useful starting skeleton; unconstrained learned refinement remained unstable.	K_5, K_6, K_13
159	Sparse and mixed audio routes	AST audio-only classification	Aligned audio supported classification and spatial selection; direct token/waveform generation remained weak.	K_1, K_3, K_4, K_13
160	Sparse and mixed audio routes	RGB-plus-audio source box	Aligned audio supported classification and spatial selection; direct token/waveform generation remained weak.	K_1, K_3, K_4, K_13
161	Sparse and mixed audio routes	EnCodec visual-to-audio baseline	Aligned audio supported classification and spatial selection; direct token/waveform generation remained weak.	K_1, K_3, K_4, K_13
162	Sparse and mixed audio routes	Image-prefix MusicGen	Aligned audio supported classification and spatial selection; direct token/waveform generation remained weak.	K_1, K_3, K_4, K_13
163	Sparse and mixed audio routes	Dense audio-semantic mask	Aligned audio supported classification and spatial selection; direct token/waveform generation remained weak.	K_1, K_3, K_4, K_13
164	Sparse and mixed audio routes	RGB-plus-audio region selector	Aligned audio supported classification and spatial selection; direct token/waveform generation remained weak.	K_1, K_3, K_4, K_13
165	Dense categorical-mask pilots	Gated fixed-class route	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
166	Dense categorical-mask pilots	Balanced-flow route	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
167	Dense categorical-mask pilots	Self-contained prototype head	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
168	Dense categorical-mask pilots	Set-level multi-label head	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
169	Dense categorical-mask pilots	Predicted-mask multi-label head	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
170	Dense categorical-mask pilots	Pixelwise semantic prototype CE	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
171	Dense categorical-mask pilots	Prototype CE with multi-label classifier	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10
172	Dense categorical-mask pilots	Fresh prototype CE without classifier	Full generation exposed background collapse and fragmented legal colors hidden by partial previews.	K_5, K_8, K_10

Continued on the next page.

Table 21 continued.

ID	Family	Variant	Direct observation and knowledge retained	Formal link
173	Categorical objective modes	Class-balanced CE	Better categorical losses improved candidates but continuous generation still needed symbolic recovery.	K_8, K_10, K_13
174	Categorical objective modes	Combined categorical objective	Better categorical losses improved candidates but continuous generation still needed symbolic recovery.	K_8, K_10, K_13
175	Categorical objective modes	Global-margin supervision	Better categorical losses improved candidates but continuous generation still needed symbolic recovery.	K_8, K_10, K_13
176	Categorical objective modes	Prototype CE	Better categorical losses improved candidates but continuous generation still needed symbolic recovery.	K_8, K_10, K_13
177	Categorical objective modes	Prototype CE with focal regression	Better categorical losses improved candidates but continuous generation still needed symbolic recovery.	K_8, K_10, K_13
178	Categorical region refiners	Per-Head categorical refiner	A second stage preserved spatial support while recovering exact labels, with fine-grained limits.	K_8, K_10, K_13
179	Categorical region refiners	Multi-condition categorical refiner	A second stage preserved spatial support while recovering exact labels, with fine-grained limits.	K_8, K_10, K_13
180	Categorical region refiners	Foreground recovery	A second stage preserved spatial support while recovering exact labels, with fine-grained limits.	K_8, K_10, K_13
181	Categorical region refiners	Coarse-label remapping	A second stage preserved spatial support while recovering exact labels, with fine-grained limits.	K_8, K_10, K_13
182	Universal Pose custom Head	Full-frame Universal Pose	Correct geometry, cropping, and gradient supervision mattered more than simply enlarging the model.	K_1, K_5, K_13
183	Universal Pose custom Head	Orientation-corrected letterbox	Correct geometry, cropping, and gradient supervision mattered more than simply enlarging the model.	K_1, K_5, K_13
184	Universal Pose custom Head	Instance-crop Universal Pose	Correct geometry, cropping, and gradient supervision mattered more than simply enlarging the model.	K_1, K_5, K_13
185	Universal Pose custom Head	Gradient-supervised Universal Pose	Correct geometry, cropping, and gradient supervision mattered more than simply enlarging the model.	K_1, K_5, K_13
186	Universal Pose custom Head	Focal-supervised Universal Pose	Correct geometry, cropping, and gradient supervision mattered more than simply enlarging the model.	K_1, K_5, K_13
187	Pose-detail objective pilots	Latent endpoint continuation	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
188	Pose-detail objective pilots	Decoded endpoint continuation	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
189	Pose-detail objective pilots	Topology continuation	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
190	Pose-detail objective pilots	Low-noise detail anchor	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
191	Pose-detail objective pilots	Soft-distance continuation	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
192	Pose-detail objective pilots	Multi-scale-gradient continuation	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
193	Pose-detail objective pilots	Gradient-direction continuation	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
194	Pose-detail objective pilots	Hard-foreground mining	Endpoint-oriented and detail losses often removed fingers; the gradient baseline remained strongest.	K_5, K_8, K_13
195	Named custom-variable routes	MAGICK transparency route	Named graph variables expanded the task graph without consuming fixed WSG channel IDs.	K_1, K_7, K_13
196	Named custom-variable routes	IC-Light relighting route	Named graph variables expanded the task graph without consuming fixed WSG channel IDs.	K_1, K_7, K_13
197	Runtime route composition	Text-to-RGB then RGB-to-world chain	Installed routes executed as chains without joint retraining while retaining edge provenance.	K_7, K_14
198	Runtime route composition	Generic intermediate-node route composition	Installed routes executed as chains without joint retraining while retaining edge provenance.	K_7, K_14

14.12. From Four Initial Tunnels to 1,148 Trained Baselines

The order of these experiments is important. We first studied four representative TunnelingNets, Heads 003, 270, 275, and 313, covering the four interfaces. Together with the earlier bridge, codec, sparse-representation, and routing experiments, these four Tunnels revealed one common rule: keep the shared generator fixed, describe the known information as the condition, describe the requested information as the target, and train the corresponding TunnelingNet to redirect generation toward that target.

The development record also compares harder and easier designs for the same task. The rule requires the chosen representation to preserve the distinctions needed by the task and to present them in a form the generator can use. Early sparse-codec experiments showed that empty background could overwhelm isolated points. A one-shot dense rendering of keypoints or boxes therefore risked losing locations, identities, or connections. The later single-canvas, semantic-color, and pseudocolor-box experiments were harder to optimize than the revised designs and lost exactly these distinctions. Multilane heatmaps, fixed person–keypoint slots, and native structured outputs preserved the missing distinctions and made the same tasks markedly easier to learn.

The categorical experiments showed a second case. Continuous generation could draw the spatial support of a region, but an approximate color could not reliably identify an exact class. Eight dense categorical pilots produced background collapse, mixed legal colors, or fragmented regions. Separating generated geometry from a lightweight categorical refiner then recovered the exact labels. Universal Pose supplied a third case: in a full-frame one-shot pose image, the sparse body signal is dominated by background and small joints can disappear. The full-frame route collapsed toward black, whereas corrected geometry, instance crops, and gradient supervision preserved the body structure. The tasks themselves were possible; their first designs had discarded information that the tasks required. Once that information was preserved, the same tasks became markedly easier to learn. The earlier versions might also learn with more data, parameters, or optimization; the comparison concerns their relative difficulty. The underlying sparse-codec experiments are described in Sec. 14.2; the later heatmap, categorical, structured-output, and Universal-Pose experiments appear in Sec. 14.6. Their family-level record is given in Table 18, and the individual interventions are indexed in Table 21: variants 136–142 record the heatmap representations, 165–181 record the categorical pilots and refiners, and 182–194 record the Universal-Pose and pose-detail experiments. Each row states the implemented change, direct observation, and formal lesson.

We then used this rule to fix the interface and representation of the remaining routes, without changing the shared backbone or previously learned tasks. We subsequently trained and evaluated the other 1,144 tasks. Together with the original four, this produced 1,148 trained and evaluated visual tasks across 39 output groups, which serve as baselines for directed visual tasks that mostly had none. The development and training records show this order: the rule came first, and the 1,144 routes were trained afterward.

We also added two capabilities introduced by other papers as ordinary TunnelingNets: LayerDiffuse-style transparency [67] and IC-Light relighting [68]. Transparency used 1,000 updates, and relighting used 2,800. Section 10.3 shows the complete experiments and outputs.

14.13. Summary of the Unified Law

The two-year experimental study spans **60 architectural and methodological families** and at least **198 explicitly named model and training variants**. Table 18 attaches a concrete experimental lesson to every family, while Table 19 describes **52 representative variants** into their motivation, exact intervention, observed behavior, and causal conclusion. Altogether, we accumulated **at least 2,293 training attempts documented in surviving records**. Among them, **1,546 different training runs** retain complete configurations, and at least **1,380 trained models** retain their weights. These conservative counts exclude older experiments that were deleted, moved, or archived elsewhere. From these experiments, we obtained the following **eight principles**.

1. If the representation erases a distinction, the model cannot learn that distinction later. The sparse-pixel, semantic-color, shared-VAE, and object-structure experiments looked different, yet each representation had collapsed two states that the requested output needed to distinguish. No architecture or loss can recover a joint identity, object membership, or physical value after the boundary has erased it. Heatmaps, fixed slots, typed sequences, object atoms, and named custom fields became learnable when each required distinction received a stable address and a faithful encoding. The first test for a new capability is thus whether its information survives both entry into and exit from the model.

2. Two latent spaces can have the same shape and distribution while speaking different internal languages. Dense fields, sparse structures, audio, language, and RGB can describe one world state while assigning different meanings to their coordinates. Whitening, alignment losses, invertible transforms, and the shadow-variable experiments could match dimensions or global statistics while decoded samples remained wrong. Mutual readability requires three things: a faithful codec for each side, stable identities for the information being exchanged, and paired sample-level relations that teach what one coordinate system means to

the other.

3. As a representation moves from detailed visual information toward abstract symbols, the useful learning objective moves from reconstruction toward prediction. An image keeps concrete location, color, geometry, and texture; a sentence keeps a much more abstract semantic account of the same scene. Moving from the first toward the second is therefore also a move from MSE-like reconstruction toward CE-like prediction of symbols. We observed the same difference among visual latents. A compact, low-channel SD-style VAE latent has already discarded much of the spatial detail and often behaves more like a prediction target. A high-channel FLUX or RAE [87] representation retains more visual structure and is better suited to masking, aligned translation, and conditional generation. The VAE acts as a visual tokenizer: details removed by the codec must be supplied by the generator, while details preserved by the codec can be reconstructed or transformed directly. Channel count alone is not the rule; what information the representation keeps is the rule.

4. Reversing condition and target changes what information is missing and can turn an easy estimate into a hard generation problem. RGB often constrains depth strongly, while depth does not determine color, texture, illumination, identity, or background. RGB-to-depth can therefore behave like estimation; depth-to-RGB must choose among many valid appearances. The missing choices must come from text, another condition, sampled noise, or a pretrained target-domain prior. Reversing the same pair of fields creates a new learning problem, so a visual task is defined by its condition, target, and direction.

5. A shorter or smoother path cannot teach a model what a valid target looks like. Optimal transport shortened and uncrossed reference segments; reflow, intermediate supervision, repeated correction, and noise conditioning improved trajectory geometry or robustness. None of them supplied textures, shapes, or other target semantics absent from the bridge. Reliable conversion emerged when a pretrained target generator already knew the target distribution and a small learned path redirected that generator's native trajectory. Path design controls how the model travels; the target prior determines what valid destinations look like.

6. Each new condition–target relation should be learned separately and then installed inside the shared model. Joint video lanes, fused conditions, packed LoRAs, and shared-target LoRAs coupled the memory, gradients, sampling schedule, and future revision of otherwise different relations. Adding or repairing one relation could then disturb the others. Separately acquired tunnels kept the generative backbone shared while storing each condition–target relation in its own small path; inference activates only the path requested by the current task. Capability addition becomes a local installation, while the installed collection still executes as one model.

7. Thousands of tasks become practical only when each new relation is cheap and leaves earlier relations untouched. Full-codec retraining changed a shared representation; external bridges had to learn a target generator; joint multimodal generators coupled every relation; packed LoRAs still repeated target-side computation and gradients. Residual interventions inside a mature generator changed only a small task-relevant path and left installed paths untouched. This bounded marginal parameter, training, and revision cost is what made 1,148 routes practical; a method whose new task reopens all old tasks cannot grow at the same rate.

8. A training loss or intermediate preview is not success; the complete inference process must produce a valid final target. Partial denoising previews, low latent losses, private round trips, and aligned global statistics sometimes looked successful while complete inference produced black images, fragmented categories, missing structures, or codes unreadable by the native decoder. We therefore run the full trajectory, decode it through the target-native interface, and evaluate the resulting object with task-native readers and metrics. Intermediate diagnostics explain failure; they do not by themselves establish a usable capability.

These eight principles supplied the experimental foundation of the Unified Law. Their common structure is direct: known RGB, text, and world information act as conditions, and requested information is generated as the target. Estimation, control, reconstruction, inverse rendering, structured prediction, and cross-property translation differ in the placement and direction of those variables.

Beginning with the initial WSG idea in 2024, the next two years of development covered **60 experimental families** and **198 named variants**. Across this process, we accumulated **at least 2,293 training attempts documented in surviving records**. Complete configurations remain for 1,546 different training runs, and weights remain for at least **1,380 trained models**. We repeatedly changed representations, objectives, architectures, training paths, and condition–target directions, then compared which behaviors survived those changes. The recurring results gradually became the eight principles above. Once this evidence converged, we formally stated the Unified Law and used it to develop the complete framework: WSG organizes RGB and additional world information, WSM installs the resulting visual capabilities around a shared generative backbone, and TunnelingNet redirects that backbone toward each requested target at low marginal cost. Using this final framework, we then trained and evaluated **1,148 visual tasks** across 39 output groups and all four interfaces. The two-year experimental process and the final evaluation are therefore consecutive stages of one study: the first explains how the law and mechanism were discovered, while the sec-

ond shows how broadly the resulting method works. These numbers describe different stages: the documented training attempts show how we developed the method, and the 1,148 tasks measure how broadly the completed method was evaluated.