HYDRA-X: Native Unified Multimodal Models with Holistic Visual Tokenizers

Guozhen Zhang[1,†,*] **Xuerui Qiu**[2,4,†,*] **Yutao Cui**[3,†] **Tianhui Song**[3] **Changlin Li**[3]
Junzhe Li[3] **Tao Huang**[3] **Xiao Zhang**[3] **Yang Li**[3] **Jianbing Wu**[3]
Miles Yang[3] **Zhao Zhong**[3] **Liefeng Bo**[3] **Limin Wang**[1,5,‡]

[1]Nanjing University [2]Institute of Automation, Chinese Academy of Sciences
[3]Tencent Hunyuan [4]Zhongguancun Academy [5]Shanghai AI Lab
zgzaacm@gmail.com lmwang@nju.edu.cn

Abstract

Holistic visual tokenizers are fundamental to unified multimodal models (UMMs) as they map diverse visual inputs into a unified representation space. In this paper, we present HYDRA-X, the first UMM framework that unifies image and video tokenization within a single Vision Transformer (ViT). Our design is driven by two core challenges: efficiently injecting spatiotemporal reconstruction capability into a native ViT, and embedding image- and video-level semantic awareness into the latent space. To address the first, comprehensive ablations reveal two key findings: **(1)** frame-level causal temporal attention suffices for visual reconstruction, whereas full spatiotemporal attention degrades it; and **(2)** hierarchical temporal compression substantially outperforms single-step alternatives. To tackle the second, we propose a lightweight decompressor that upsamples temporally compressed features under joint image-video teacher supervision, thereby enforcing complementary semantic structures within the compact latent space. Building on this holistic tokenizer, we further propose a principled improvement of the editing pipeline: source-target interaction should occur at the latent level inside the tokenizer rather than at the semantic level inside the LLM, substantially improving editing consistency and accelerating convergence. Instantiated at the 7B dense model, HYDRA-X achieves strong performance across image and video understanding and generation tasks, laying a solid foundation for future unified-tokenizer UMM exploration.

1 Introduction

Native unified multimodal models (UMMs) [1–5] have recently emerged as a powerful paradigm that jointly trains a single autoregressive backbone for both visual understanding and generation. A central design choice is how visual inputs are encoded: existing systems either deploy *decoupled visual encoders* that pair a ViT encoder with a separate VAE encoder for the two tasks [3, 5, 6], or adopt a *unified visual tokenizer* that maps diverse visual inputs into a single representation space shared by both tasks [1, 2, 4, 7–9]. The latter approach offers distinct architectural advantages: it eliminates the representational mismatch between heterogeneous encoders that the LLM must otherwise reconcile, and enables mutual reinforcement between understanding and generation.

While unified tokenization has been extensively explored for static images, a *holistic* tokenizer that binds images and videos into a single representation space has received much less attention. Existing video-capable UMMs typically adopt one of two ad-hoc strategies. The first paradigm relies on frame-wise tokenizers that apply an image semantic encoder independently to

*Work done during internship at Tencent Hunyuan. † Equal contribution. ‡ Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

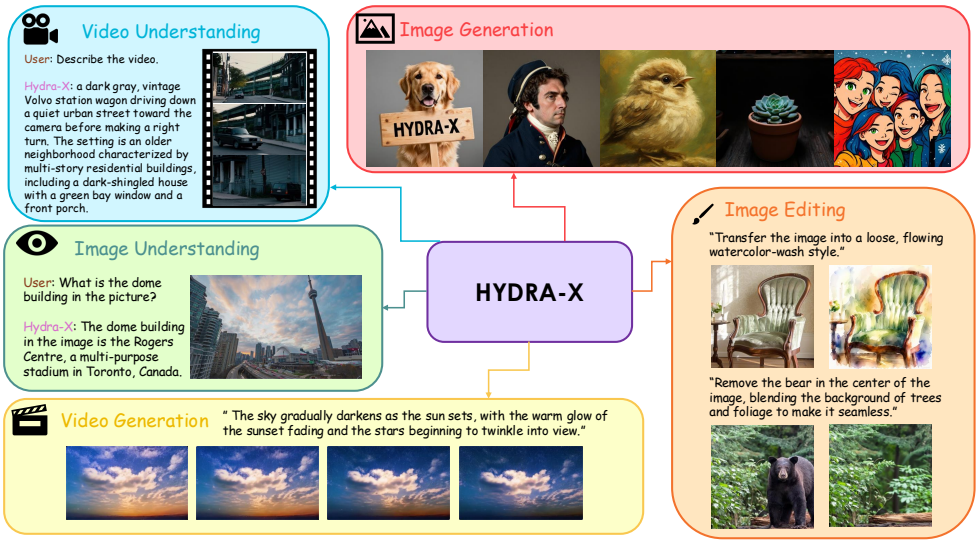


Figure 1: HYDRA-X is a native UMM that unifies image/video understanding, image/video generation, and instruction-guided image editing through one holistic tokenizer HYDRA-XTOK.

each frame [7]. Without any temporal interaction inside the tokenizer, the resulting representation cannot capture cross-frame dynamics such as motion or short-horizon causality, leaving the downstream LLM with disjoint per-frame features that carry no inherent video structure. The second paradigm employs cascaded designs that stack a 3D causal VAE before a semantic encoder [1, 2]. Although this packs the temporal axis, the VAE is trained in isolation without any semantic constraint, and may inadvertently discard information critical for understanding.

In this paper, we present HYDRA-X, the first UMM framework built upon HYDRA-XTOK, a unified visual tokenizer that handles both image and video encoding within a single Vision Transformer (ViT). Our overall design follows the image-only UMM framework HYDRA [4], which compresses intermediate-layer ViT features into a compact latent and then reconstructs semantic feature from it. Extending this paradigm to jointly support images and videos, however, raises two core challenges: *(a)* efficiently injecting spatiotemporal reconstruction capability into a native ViT, and *(b)* embedding both image- and video-level semantic awareness into the shared latent space.

Our investigation of the first challenge yields two findings that run counter to conventional intuition. **(1)** Although full spatiotemporal attention is the most natural extension to video, it actively degrades reconstruction by disrupting the locality and structure encoded during image pretraining. Surprisingly, frame-level causal temporal attention with a minimal temporal receptive field, attending only to the immediately preceding frame, comprehensively outperforms its global counterpart. **(2)** A single-step patchify substantially underperforms a hierarchical patchify that distributes temporal compression across multiple stages, indicating that the temporal axis benefits from progressive, multi-scale folding. Together, these two design choices enable HYDRA-XTOK to surpass the reconstruction fidelity of dedicated 3D-conv video VAEs such as Wan2.2-VAE [10].

To address the second challenge, we extend the established paradigm of semantic distillation [4, 8, 9] from images to video, and uncover a fundamental asymmetry: while image latents can readily reuse existing semantic teachers, no available video encoder operates at the compressed temporal resolution of our latent, leaving the video stream without a natural source of semantic supervision. We resolve this asymmetry through a remarkably simple addition: a lightweight *Decompressor* that lifts the compressed latent back to its native temporal length, enabling direct distillation from pretrained image and video teachers [11, 12] at full frame rate. Under this dual spatiotemporal supervision, the compact latent simultaneously preserves pixel-level fidelity and rich spatiotemporal semantic structure, substantially advancing both understanding and generation in UMMs.

Building on this holistic tokenizer, HYDRA-X unifies five UMM tasks within a single shared encoder, as shown in Figure 1: image/video generation, image/video understanding, and image editing. Yet editing in particular exposes a fundamental flaw in both HYDRA and cascaded designs: by feeding the LLM only post-encoder semantic features, they confine source-target interaction to the

semantic level and forfeit the fine-grained structural information that resides at the latent. To resolve this, we propose a principled inversion of the design: HYDRA-XTOK jointly tokenizes source and target with cross-frame interaction, fusing structural details directly into the target before reaching the LLM. This early interaction improves editing consistency and accelerates convergence.

Instantiated at the 7B scale on top of Qwen2.5-7B-Instruct [13], HYDRA-X achieves strong performance across image and video understanding and generation tasks. More importantly, it elevates the visual tokenizer from a specialized image-processing component to a holistic image-and-video interface, laying a solid foundation for future unified-tokenizer UMM exploration.

2 Related Work

Visual tokenizers for UMMs. A growing body of work unifies reconstruction and semantics within a single visual tokenizer. For images, RAE [14, 15] freezes a semantic encoder and learns a pixel decoder, while several unified-tokenizer designs [9, 16–21] co-train reconstruction and understanding within a single ViT. HYDRA [4] introduces a progressive ViT with a Generation-Semantic Bottleneck for compress-then-restore semantic distillation, which HYDRA-XTOK inherits. Aligning generative latents with semantic features has further been shown to mutually benefit both tasks [22–26]. Joint image-and-video tokenization, however, remains largely under-explored: for video, 3D-convolutional VAEs [10, 27] dominate but lack any semantic structure. A recent work, AToken [28], unifies images and videos within a single tokenizer for reconstruction and understanding, but emits task-specific output features for the two objectives and therefore does not yield the unified representation. HYDRA-X is, to our knowledge, the first UMM framework to unify image and video within a single ViT-based tokenizer, augmenting HYDRA's philosophy with explicit temporal causality, hierarchical patchify, and a Decompressor for spatiotemporal semantic awareness.

Native unified multimodal models. UMMs handle visual understanding and generation within a single backbone. Composite UMMs [21, 29–32] connect pretrained understanding and generation models via adapters, leaving the synergy between the two tasks shallow. Native UMMs instead train both objectives jointly: quantised-token approaches [5, 33–36] suffer from VQ-induced reconstruction loss, while decoupled designs [3, 6, 37–41] adopt separate semantic and VAE encoders that trade representational competition for parameter redundancy. Most relevant to our work are unified-encoder UMMs such as TransNext [7], Show-o2 [1], and TUNA [2], which share a single visual tokenizer across both tasks. We extend this line in two directions: a unified ViT tokenizer for images and videos, and tokenizer-stage source-target interaction for editing.

3 Preliminaries: Representation-Harmonized Tokenization

Our overall design follows the image-only UMM framework HYDRA [4]. At its core is a single ViT split into a *Gen-ViT* and a *Sem-ViT*, connected by a *Generation–Semantic Bottleneck* that supports generation and semantic perception within one backbone. Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, the Gen-ViT first produces a feature $\mathbf{h}$ rich in structural primitives, which the Bottleneck projects into a compact latent $\mathbf{z} \in \mathbb{R}^{N\times C}$ suitable for generation. The Sem-ViT then un-projects $\mathbf{z}$ back into a high-dimensional semantic feature $\mathbf{s}$, distilled from a pretrained semantic teacher $\mathcal{T}$:

$$\mathbf{x} \xrightarrow{\text{Gen-ViT}} \mathbf{h} \xrightarrow{\text{Bottleneck}} \mathbf{z} \xrightarrow{\text{Sem-ViT}} \mathbf{s} \xleftarrow{\text{align}} \mathcal{T}(\mathbf{x}). \tag{1}$$

The downstream LLM operates exclusively on the Sem-ViT output $\mathbf{s}$ for both understanding and generation, whereas the pixel decoder that reconstructs images from $\mathbf{z}$ is invoked only during tokenizer training. We retain this overall design and extend it from images to videos through explicit temporal causality, hierarchical patchify, and a Decompressor introduced in Section 4.

4 HYDRA-XTOK: Holistic Visual Tokenization in a Single ViT

HYDRA-XTOK is designed as the visual interface of HYDRA-X: before any token reaches the LLM, it must be compact enough for generation, faithful enough for reconstruction, and semantic enough for understanding. We initialize Gen-ViT and Sem-ViT from SigLIP 2 [11]; all UMM-side ablations use Qwen2.5-1.5B [13]. Training combines reconstruction with two semantic distillation terms:

$$\mathcal{L}_{\text{HYDRA-XTOK}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{dist}}, \tag{2}$$

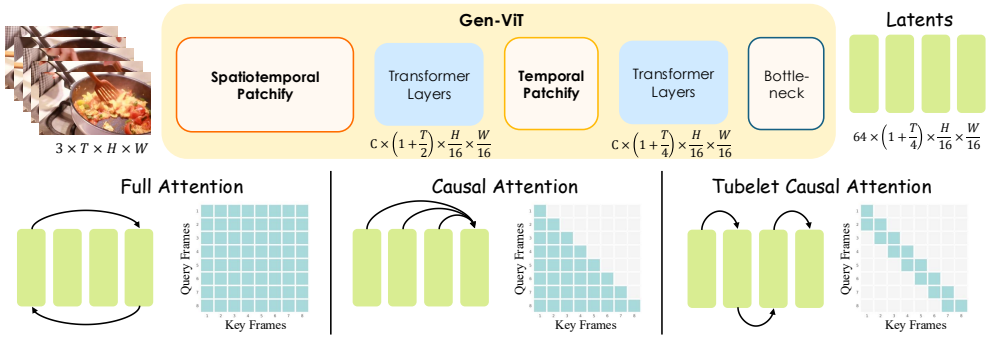


Figure 2: **Spatiotemporal reconstruction design.** (Top) The Gen-ViT folds a clip into a compact latent. (Bottom) Three ablated attention masks: Full attends across all space-time tokens, Causal masks future frames, and Tubelet restricts attention to a 2-frame window.

Table 1: **Reconstruction ablation on ImageNet** (256×256) **and DAVIS** ($17 \times 256 \times 256$)**.** All three attention-mask baselines use the single-step $4\times$ temporal patchify; only the 'Ours' row uses the hierarchical 2×2 schedule. Latency is measured per forward pass on a $17\times512\times512$ video clip. Further reconstruction comparisons and visualizations are provided in Appendix H.1 and H.3

Method	Latency (s) (↓)	ImageNet			DAVIS		
		PSNR (↑)	SSIM (↑)	rFID (↓)	PSNR (↑)	SSIM (↑)	rFVD (↓)
Full attention	0.49	31.10	0.8890	0.367	27.40	0.8277	16.20
Causal attention	0.45	31.38	0.8901	0.352	27.62	0.8283	14.05
Tubelet attention	**0.17**	31.42	0.8907	0.347	27.69	0.8287	13.69
Ours	0.25	**31.73**	**0.8936**	**0.329**	**27.97**	**0.8307**	**11.19**

where $\mathcal{L}_{\text{rec}}$ keeps the compact latent pixel-faithful, $\mathcal{L}_{\text{dist}}$ aligns Sem-ViT features with semantic features. Detailed recipes are in Appendix A.1.

4.1 Spatiotemporal Reconstruction in a ViT

Existing ViT-based tokenizers that jointly handle images and videos reconstruction, such as AToken [28] and OmniTokenizer [42], share two design choices: full spatiotemporal attention across all frames, and a single-step temporal patchify applied at the input that aggressively compresses the temporal axis. Both choices come at a cost. Full spatiotemporal attention scales quadratically with the clip length and tends to disrupt the per-frame structural prior inherited from image pretraining; the aggressive single-step patchify collapses fine-grained temporal details before any cross-frame reasoning. This naturally raises a critical question: are these design choices really necessary?

We answer this through a controlled ablation along the same two axes: *(i)* the temporal attention region, and *(ii)* the temporal patchify schedule. Following the common design in video VAEs, a clip $\mathbf{x} \in \mathbb{R}^{3\times(1+T)\times H\times W}$ is encoded into an anchor image latent together with the remaining T frames compressed by a factor of 4, producing a compact latent $\mathbf{z} \in \mathbb{R}^{C\times(1+\frac{T}{4})\times\frac{H}{16}\times\frac{W}{16}}$. The two axes are then ablated independently. For *(i)* we compare three attention masks (Fig. 2, bottom): *Full attention*, the standard choice in AToken and OmniTokenizer; *Causal attention*, with a causal mask across all preceding frames; and *Tubelet attention*, where causal attention is restricted to a 2-frame tubelet so each token attends only to its own frame and the immediately preceding one. For *(ii)* we compare the single-step $4\times$ temporal patchify used by AToken and OmniTokenizer against a hierarchical schedule that applies two consecutive $2\times$ patchify stages (top of Fig. 2). During each temporal patchify stage, the anchor frame is zero-padded so that it goes through the same operation as the remaining frames.

Table 1 reveals two principles that contradict the choices made in existing tokenizers. First, expanding the temporal receptive field beyond a 2-frame tubelet only degrades reconstruction: both full bidirectional and all-past causal attention perform worse than Tubelet attention. Second, distributing temporal compression across two patchify stages consistently outperforms a single-step

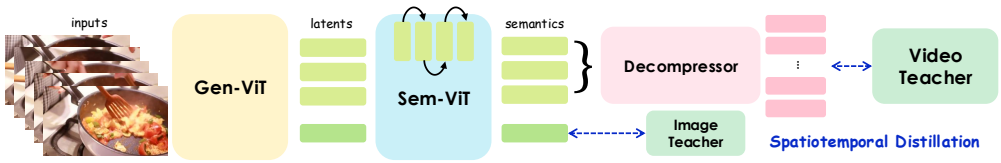


Figure 3: **Spatiotemporal distillation.** The uncompressed image latent is directly distilled by an image teacher; the $4\times$ temporally-compressed video latent is first lifted to origin length T by a lightweight Decompressor before distillation by a video teacher.

Table 2: **Semantic-distillation ablation.**

Design choices				Vid. Und.		Img. Und.		Img. Gen.	Edit
img distill	Decomp w/ img	Decomp w/ video	Sem-ViT bi-dir	MVBench (↑)	VideoMME (↑)	AI2D (↑)	MME (↑)	GenEval (↑)	ImgEdit (↑)
				29.8	27.4	45.1	989	67.5	2.35
✓				42.1	42.5	61.2	1339	70.6	2.72
✓	✓			44.7	44.3	**62.7**	**1522**	70.7	3.07
✓		✓		**45.4**	**45.0**	62.5	1501	**72.0**	**3.20**
✓		✓	✓	43.1	43.7	62.0	1434	70.1	2.70

counterpart at the same compression ratio. These results answer our opening question: aggressive spatiotemporal attention and single-step patchify are not only unnecessary but actively suboptimal.

- **Less attention is more.** Restricting causal attention to a 2-frame tubelet yields the best reconstruction among the variants we tested; wider receptive fields, including full bidirectional and all-past causal, disturb local detail more than they help.
- **Hierarchical patchify outperforms single-step.** Distributing temporal compression across two $2\times$ patchify stages consistently improves reconstruction over a single $4\times$ patchify, indicating that the temporal axis benefits from progressive, multi-scale folding.

4.2 Spatiotemporal Semantic Distillation via the Decompressor

Following HYDRA, we inject semantic structure into the latent by distilling the Sem-ViT output against pretrained teachers. Extending this recipe to video, however, reveals a fundamental asymmetry. For images, the Sem-ViT output has the same spatial resolution as a frame and can be aligned token-by-token with an off-the-shelf image teacher. For video, the Sem-ViT output is temporally compressed to $1 + T/4$ tokens, while existing video encoders operate at the original frame rate. The standard recipe therefore provides no video-level semantic supervision.

We resolve this asymmetry by introducing a lightweight *Decompressor*, a small ViT module $\mathbf{D}$ that lifts the temporally compressed Sem-ViT output back to its native temporal length, producing dense per-frame semantic features that can be aligned with both image and video teachers (Fig. 3). The Decompressor is only used at tokenizer-training time and is discarded afterwards; the LLM still operates on the same compact Sem-ViT output $\mathbf{s}$. Letting $d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos(\mathbf{a}, \mathbf{b})$ denote the cosine distance, the full distillation loss combines an image-teacher term at $\mathbf{s}$ and a video-teacher term at the Decompressor output:

$$\mathcal{L}_{\text{dist}} = d_{\cos}\big(\mathbf{s}_0,\, \mathcal{T}_{\text{img}}(\mathbf{x})\big) + d_{\cos}\big(\mathbf{D}(\mathbf{s}_{1:}),\, \mathcal{T}_{\text{vid}}(\mathbf{x})\big), \tag{3}$$

where $\mathbf{s}_0$ is the leading uncompressed image token and $\mathbf{s}_{1:}$ are the compressed video latents. For pure image batches, the video term in Eq. 3 is masked out. We ablate four design choices in Table 2: *(i)* whether to apply image distillation at the Sem-ViT output (*img distill*); *(ii)* whether to additionally distill the Decompressor output against an image teacher (*Decomp w/ img*); *(iii)* or against a video teacher (*Decomp w/ video*); and, as a cross-check of F1, *(iv)* whether the Sem-ViT uses bidirectional rather than tubelet attention (*Sem-ViT bi-dir*).

Table 2 surfaces three principles. First, semantic distillation is indispensable: removing it collapses both image and video understanding. Second, the Decompressor is what unlocks video-level supervision: distilling it against a video teacher yields the strongest video understanding while preserving

image-side performance, and the same configuration also delivers the best image generation and editing scores, consistent with the hypothesis that semantically richer latents accelerate the LLM's convergence on generation and editing. Third, switching the Sem-ViT to bidirectional attention uniformly degrades every metric, mirroring F1: less attention also benefits understanding.

- **A semantic latent lifts both understanding and generation.** Dual image- and video-teacher distillation, enabled by the Decompressor, equips the compact latent with explicit spatiotemporal semantic structure, jointly improving understanding and generation.

Discussion. The video loss supervises $\mathbf{D}(\mathbf{s}_{1:})$, whereas the LLM consumes compressed $\mathbf{s}_{1:}$. We directly probe the latter (Table 3). With the tokenizer frozen, a linear classifier on spatially and temporally mean-pooled features reaches 47.8% Kinetics-400 top-1 [43], versus 38.9% for image-teacher decompression and 0.3% without distillation. Probing the decompressed features instead gives 53.2%: action information is accessible before decompression, though the expanded representation improves linear readout. A separate linear probe predicts which of four shuffled sub-clips came first. Video supervision raises accuracy to 68.7%, compared with 39.4% for image-teacher decompression and 27.6% without distillation (chance: 25%).

As a control, we bypass $\mathbf{D}$ and directly distill from video-teacher features temporally average-pooled to the compressed resolution. This yields $41.8/42.3$ on MVBench/Video-MME and 42.7% on K400, below $45.4/45.0$ and 47.8% with the Decompressor. These probes support the role of full-resolution video supervision through $\mathbf{D}$ in shaping action and temporal information in the features.

Table 3: **Direct verification of the Decompressor mechanism.** Frozen-feature probes use compressed $\mathbf{s}_{1:}$ unless marked $\mathbf{D}(\mathbf{s}_{1:})$. K400 and temporal-order scores are accuracies (%); higher is better throughout.

(a) Action recognition

Distillation	K400
None	0.3
Image only	37.6
Decomp. + image	38.9
Decomp. + video	**47.8**
Same model, probe $\mathbf{D}(\mathbf{s}_{1:})$	53.2

(b) Temporal-order prediction

Distillation	Accuracy
None	27.6
Decomp. + image	39.4
Decomp. + video	**68.7**

Chance accuracy: 25%.

(c) Direct distillation versus temporal decompression

Video supervision	MVBench	Video-MME	K400
Direct, temporally pooled teacher	41.8	42.3	42.7
Through Decompressor	**45.4**	**45.0**	**47.8**

5 HYDRA-X: Advancing Unified Multimodal Models with Holistic Tokenizers

5.1 Overall Architecture

HYDRA-X follows the standard native UMM template [1, 2, 4]: text tokens and visual tokens produced by HYDRA-XTOK are interleaved into a single sequence and processed by a shared LLM backbone with two specialised heads, an autoregressive language head trained with next-token prediction and a vision head trained with flow matching [44, 45]. Within this template, HYDRA-X unifies five tasks under one shared tokenizer HYDRA-XTOK (Fig. 4(a)): image generation (text $\rightarrow$ image), image understanding (image $\rightarrow$ text), video generation (text $\rightarrow$ video), video understanding (video $\rightarrow$ text), and image editing (source image with text instruction $\rightarrow$ target image).

As illustrated in Fig. 4(c), the same Gen-ViT serves all five tasks; the only task-dependent component is which head decode the LLM output. The model is trained end-to-end with the composite loss

$$\mathcal{L}_{\text{HYDRA-X}} = \lambda_1 \mathcal{L}_{\text{NTP}} + \lambda_2 \mathcal{L}_{\text{FM}}, \tag{4}$$

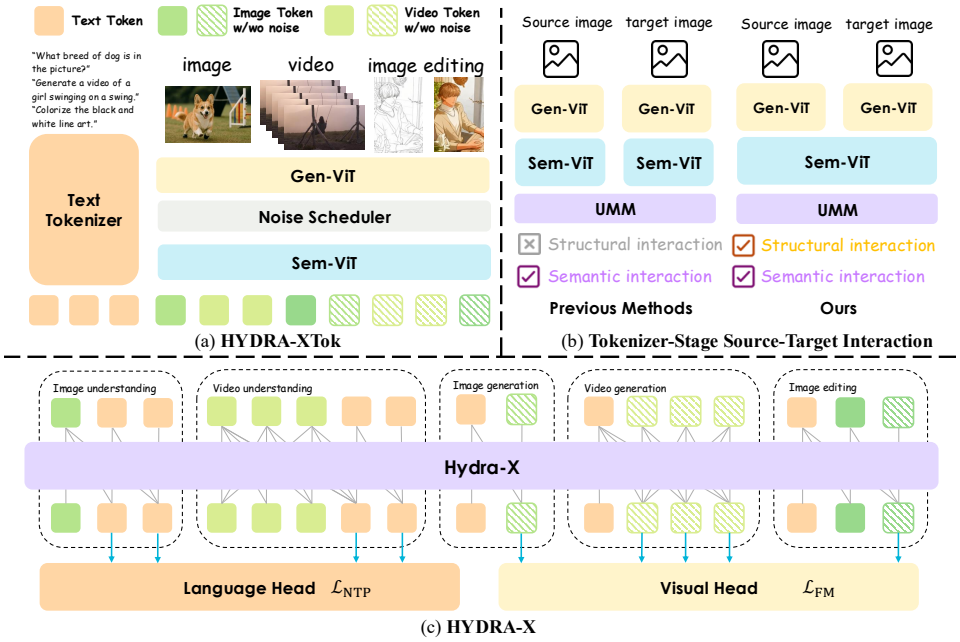


Figure 4: **HYDRA-X unifies five visual tasks through the holistic tokenizer HYDRA-XTOK. (a)** HYDRA-XTOK encodes any image or video into a compact Gen-ViT latent and then into semantic features with Sem-ViT. **(b)** Previous editing pipelines (left) encode source and target with two independent branches; HYDRA-X (right) keeps Gen-ViT independent for faithful reconstruction but shares the Sem-ViT with tubelet causal attention, injecting structural interaction inside the tokenizer. **(c)** A shared backbone with one language head and one vision head drives all five tasks.

where $\mathcal{L}_{\text{NTP}}$ is the next-token prediction loss for text, $\mathcal{L}_{\text{FM}}$ is the rectified flow matching loss for visual latents, and both loss weights λ_1 and λ_2 are set to 1 by default.

5.2 Independent Encoding Bypasses the Latent

Among the five tasks, image editing is the only one whose input contains both a conditioning image and a target image. Conventional pipelines, including HYDRA [4] and cascaded designs such as Show-o2 [1] and TUNA [2], tokenise the source $\mathbf{x}_c$ and the target $\mathbf{x}_t$ *independently* with the same tokenizer (Fig. 4(b), left):

$$[\mathbf{s}_c,\, \mathbf{s}_t] \;=\; \big[\text{HYDRA-XTOK}(\mathbf{x}_c),\, \text{HYDRA-XTOK}(\mathbf{x}_t)\big], \qquad \mathbf{s}_c \perp \mathbf{s}_t \quad \text{inside the tokenizer.} \tag{5}$$

As a result, the source and target latents $\mathbf{z}_c, \mathbf{z}_t$ never interact inside the tokenizer, and the LLM has to discover their cross-image alignment from scratch on top of two independent semantic streams. Detail-faithful editing remains challenging, particularly over multiple editing rounds [46].

5.3 Tokenizer-Stage Source-Target Interaction

A natural fix falls out of HYDRA-XTOK's holistic design: since the Sem-ViT already applies tubelet causal attention for video modeling, we reuse the exact same mechanism for editing pairs by routing $(\mathbf{x}_c, \mathbf{x}_t)$ through HYDRA-XTOK as a length-2 clip (Fig. 4(b), right). The Gen-ViT continues to encode the two images independently and the post-Bottleneck latents $\mathbf{z}_c, \mathbf{z}_t$ remain reconstruction-faithful. The cross-image interaction is then injected exclusively at the Sem-ViT, which processes $[\mathbf{z}_c; \mathbf{z}_t]$ with the same tubelet causal mask used for video:

$$[\mathbf{s}_c,\, \mathbf{s}_t] \;=\; \text{Sem-ViT}\big([\mathbf{z}_c; \mathbf{z}_t]\big), \quad \text{causal: } \mathbf{s}_c \text{ attends only to } \mathbf{z}_c, \;\; \mathbf{s}_t \text{ attends to } [\mathbf{z}_c; \mathbf{z}_t]. \tag{6}$$

Note that for editing pairs we disable Gen-ViT's cross-frame tubelet attention since the source and target are not temporally adjacent video frames; only Sem-ViT (the semantic stage) reuses the video tubelet causal mask. This asymmetric reuse is a deliberate choice: structural reconstruction benefits from independent encoding, while semantic alignment benefits from cross-image interaction.

Table 4: **Source-target interaction (STI) ablation.** HYDRA-X-STI tokenises the editing pair as a length-2 clip with Sem-ViT tubelet causal attention; HYDRA-X-Indep encodes the source and target independently. *Recon-PSNR*: PSNR of source reconstruction on ImgEdit.

Layout	Editing		Img. Gen.	Vid. Und.		Img. Und.	
	ImgEdit (↑)	Recon-PSNR (↑)	GenEval (↑)	MVBench (↑)	VideoMME (↑)	AI2D (↑)	MME (↑)
HYDRA-X-Indep	2.80	20.74	70.51	45.3	45.0	62.2	1478.5
HYDRA-X-STI	**3.20**	**27.65**	**71.97**	**45.4**	**45.0**	**62.5**	**1501.0**

Table 4 compares HYDRA-X-Indep against HYDRA-X-STI, identical except for whether the editing pair is encoded independently or as a length-2 clip with Sem-ViT tubelet causal attention. STI raises *Recon-PSNR*, the PSNR of source reconstruction on ImgEdit [47] that directly probes editing consistency, by nearly 7 dB and lifts ImgEdit by 0.4. STI further yields consistent gains on most non-editing benchmark, with GenEval ($+1.46$) the most prominent, suggesting that the new latent-level coupling also enriches the Sem-ViT for generation. The Recon-PSNR jump directly validates our hypothesis from Section 5.2: editing's consistency failure stems from latent-level isolation inside the tokenizer, not from LLM capacity or supervision.

- **Latent-level interaction matters for editing.** Reusing the Sem-ViT's tubelet causal attention to fuse source and target inside the tokenizer, adds no parameters and no separate cross-image module yet substantially improves editing consistency.

6 Main Results

Implementation. HYDRA-X is instantiated at two scales. The reported model uses Qwen2.5 -7B-Instruct [13] as the LLM backbone; a matched 1.5B variant is used for the methodological ablations in Sections 4–5. Following AToken [28], HYDRA-XTOK includes a symmetric ViT encoder/decoder pair augmented with 3D rotary position embeddings (3D RoPE) [48] for joint spatiotemporal modelling. The Decompressor $\mathbf{D}$ in Eq. 3 is a lightweight $4\times$ temporal upsampler that stacks two consecutive *(temporal upsample $\rightarrow$ transformer block)* stages; each temporal upsample is a 1×1 convolution doubling the channel dimension ($C \rightarrow 2C$) followed by a channel-to-time reshape, inverting the encoder's hierarchical 2×2 temporal patchify. The bottleneck dimension is $C=64$. For distillation teachers, we use SigLIP-SO400M-patch16-naflex [11] as the image teacher $\mathcal{T}_{\text{img}}$ and InternVideo-Next-L [49] as the video teacher $\mathcal{T}_{\text{vid}}$.

6.1 Multimodal Understanding

Image understanding. We benchmark on AI2D [50], MME [51], MMMU [52], OCRBench [53], MMBench [54], RealWorldQA, ChartQA [55], DocVQA [56], and InfoVQA [57]. Table 5 compares HYDRA-X against open-source UMMs at a similar scale. Overall, HYDRA-X matches or exceeds 7B native UMM baselines on most reported metrics. Compared with HYDRA, it improves AI2D from 85.1 to 86.5, MME from 2068.6 to 2350.0, and OCRBench from 57.7 to 84.5, while also achieving 86.5 on ChartQA. These results demonstrate strong diagram, text, and chart understanding within a model that jointly supports video and generation. Performance on DocVQA and InfoVQA remains below Qwen2.5-VL-7B, highlighting room for improvement in document understanding.

Video understanding. We evaluate on MVBench [71], Video-MME [72], LongVideoBench [73], and LVBench [74] (Table 6). HYDRA-X improves over the reported unified baselines wherever comparable results are available. Relative to Show-o2-7B, it gains 3.3, 2.6, and 4.0 points on MVBench, Video-MME, and LongVideoBench, respectively, and achieves 30.0 on LVBench. It remains below the strongest dedicated or proprietary video LMMs on several metrics, while using one ViT tokenizer across understanding, generation, and editing. These results are consistent with the benefit of dual-teacher distillation in HYDRA-XTOK. Since training recipes differ across systems, the controlled ablations and direct feature probes in Section 4.2 provide more specific evidence that video supervision enriches the compressed representation.

Table 5: **Evaluation on image understanding benchmarks.** # Params. denotes the model size. Rows in gray indicate models with $\geq$ 14B parameters and are excluded from the ranking. Within each subgroup of the table, **bold** marks the best result and underline marks the second-best.

Models	# Params	AI2D	MME	MMMU	OCRBench	MMB	RWQA	ChartQA	DocVQA	InfoVQA
		test	summary	val	test	dev_en	test	test	val	val
Understanding-only Models										
Qwen2.5-VL [58]	7B	**84.3**	**2312.0**	**58.0**	**88.8**	**82.8**	**68.4**	**84.1**	**93.0**	**78.6**
LLaVA-1.5 [59]	7B	55.5	1510.7	35.7	31.8	62.3	54.8	17.9	–	–
LLaVA-OV [60]	7B	81.4	1998.1	48.8	62.2	80.8	66.3	80.0	87.5	68.8
Unified Multimodal Models										
BLIP3-o [30]	8B	–	2329.7	50.6	83.1	83.5	69.0	78.0	–	–
TokenFlow-XL [18]	14B	–	1922.1	43.2	–	68.9	56.6	–	–	–
Ming-UniVision [61]	16B	82.8	2023.0	40.3	72.4	78.5	–	–	–	–
BAGEL [3]	14B	89.2	2388.0	55.3	73.3	85.0	72.8	78.5	–	–
Janus-Pro [38]	7B	71.3	1567.1	41.0	59.0	79.2	58.0	25.8	–	–
VILA-U [62]	7B	–	1401.8	–	–	66.6	–	–	–	–
Show-o2 [1]	7B	78.6	1620.5	48.9	32.4	79.3	64.7	52.3	77.3	45.8
HYDRA [4]	7B	85.1	2068.6	49.4	57.7	82.4	64.7	–	–	–
HYDRA-X	7B	**86.5**	**2350.0**	**51.5**	**84.5**	**84.0**	**68.7**	**86.5**	**81.7**	**59.1**

Table 6: **Evaluation on video understanding benchmarks.** # Params. denotes the model size. Video-MME reports the w/o-subtitle score.

Models	# Params	MVBench	Video-MME	LongVideoBench	LVBench
		test	w/o sub	val	test
Proprietary Und.-only Models					
GPT-4V [63]	–	43.5	59.9	61.3	–
GPT-4o [64]	–	–	71.9	**66.7**	**48.9**
Gemini-1.5-Flash [65]	–	–	70.3	61.6	–
Gemini-1.5-Pro [65]	–	**54.2**	**75.0**	64.0	33.1
Open-source Und.-only Models					
VILA [66]	40B	–	**60.1**	–	–
PLLaVA [67]	34B	58.1	–	53.2	–
LongVA [68]	7B	49.2	52.6	51.8	–
VideoLLaMA2 [69]	7B	54.6	47.9	–	–
LLaVA-OV [60]	7B	56.7	58.2	**56.5**	**26.9**
IXC-2.5 [70]	7B	**69.1**	55.8	–	–
Unified Multimodal Models					
Show-o2 [1]	1.5B	49.8	48.0	49.2	–
TUNA [2]	1.5B	54.4	49.1	49.7	27.4
Show-o2 [1]	7B	55.8	57.4	55.5	–
HYDRA-X	7B	**59.1**	**60.0**	**59.5**	**30.0**

6.2 Visual Generation

Table 7 reports image generation on GenEval [75] and WISE [76], and video generation on VBench [77] for 17-frame outputs at 640×384, measured by Quality Score (QS), Semantic Score (SS), and Total. HYDRA-X leads every reported GenEval and WISE column among 7B-scale unified baselines and remains competitive with $\geq$ 14B unified models. It achieves Overall scores of 0.88 on GenEval and 0.56 on WISE, versus 0.76 and 0.44 for Show-o2-7B. Improvements across object composition, spatial relationships, and knowledge-oriented generation suggest that the shared representation supports both instruction adherence and broader semantic requirements.

On VBench, HYDRA-X leads all reported unified entries on QS, SS, and Total, exceeding the closest unified competitor (Show-o2-1.5B) by $+1.87$, $+3.26$, and $+2.15$, respectively. Gains in both QS and SS indicate that stronger text alignment accompanies better aggregate visual quality. Appendix Table 14 provides per-dimension results, with HYDRA-X also leading the reported unified baselines on Object Class, Human Action, and Scene. These results demonstrate the shared tokenizer's versatility across image and short-video generation. Together with the controlled ablations in Section 4.2, they support the view that video-teacher supervision enriches the latent's semantic structure while preserving its utility for visual synthesis.

Table 7: **Comprehensive visual generation results.** Image generation on GenEval [75] and WISE [76]; video generation on VBench [77] reporting Quality Score (**QS**), Semantic Score (**SS**), and the aggregate **Total** score. † refers to using LLM rewriters. Rows in gray indicate models with $\geq$ 14B parameters and are excluded from the ranking. Qualitative results are in Appendix H.4.

Models	# Params	GenEval			WISE			VBench		
		Two Obj.	Pos.	Over.	Cult.	Space	Over.	QS	SS	Total
Generation-only Models										
SD3-Med [45]	2B	**0.94**	0.33	0.74	–	–	–	–	–	–
FLUX.1 [Dev] † [78]	12B	0.93	**0.68**	**0.82**	**0.48**	**0.62**	**0.50**	–	–	–
CogVideoX [79]	5B	–	–	–	–	–	–	82.75	**77.04**	81.61
Hunyuan Video [80]	13B	–	–	–	–	–	–	**85.07**	76.88	**83.43**
Unified Multimodal Models										
TokenFlow-XL [18]	14B	0.60	0.16	0.55	–	–	–	–	–	–
SEED-X [81]	17B	0.58	0.19	0.49	–	–	–	–	–	–
Ming-UniVision [61]	16B	0.93	0.92	0.85	–	–	–	–	–	–
BAGEL † [3]	14B	0.95	0.78	0.88	0.44	0.68	0.52	–	–	–
Show-o2 [1]	1.5B	0.86	0.46	0.73	0.33	0.53	0.39	82.10	78.31	81.34
Harmon [8]	1.5B	0.86	0.74	0.76	0.38	0.52	0.41	–	–	–
Blip3-o† [30]	8B	–	–	0.84	–	–	0.50	–	–	–
MUSE-VL [82]	7B	0.64	0.25	0.57	–	–	–	–	–	–
Janus-Pro [38]	7B	0.89	0.79	0.80	0.30	0.49	0.35	–	–	–
VILA-U [62]	7B	–	–	–	0.26	0.37	0.31	76.26	65.04	74.01
HaploOmni [83]	7B	–	–	–	–	–	–	–	–	78.10
Emu3 [84]	8B	–	–	0.66	–	–	–	–	–	80.96
Show-o2 † [1]	7B	0.87	0.52	0.76	0.40	0.58	0.44	–	–	–
HYDRA-X	7B	**0.95**	**0.84**	**0.88**	**0.53**	**0.72**	**0.56**	**83.97**	**81.57**	**83.49**

6.3 Image Editing

Table 8: **Image editing. Ext.**: Extract; **Rm.**: Remove; **Over.**: mean across nine ImgEdit categories. GEdit-Bench reports semantic consistency (**G-SC**), perceptual quality (**G-PQ**), and overall score (**G-Over.**). Full results appear in Appendix Table 13.

Models	ImgEdit-Bench			GEdit-Bench		
	Ext.	Rm.	Over.	G-SC	G-PQ	G-Over.
Generation-only Models						
FLUX.1 Kontext [Pro] [78]	2.35	3.57	4.00	7.02	7.60	6.56
Qwen-Image [85]	**3.43**	**4.14**	**4.27**	**8.00**	**7.86**	**7.56**
Unified Multimodal Models						
OmniGen [86]	1.71	2.43	2.96	5.96	5.89	5.06
UniWorld-V1 [31]	2.27	3.24	3.26	4.93	**7.43**	4.85
BAGEL [3]	1.70	2.62	3.20	7.36	6.83	6.52
OmniGen2 [87]	1.77	3.20	3.44	7.16	6.77	6.41
HYDRA-X	**4.04**	**4.38**	**4.34**	**7.80**	7.24	**7.17**

Table 8 reports results on GEdit-Bench [88] and ImgEdit-Bench [47]. Among 7B-scale UMMs, HYDRA-X leads on **Ext.** (4.04, +1.77), **Rm.** (4.38, +1.14), ImgEdit **Over.** (4.34, +0.90), and GEdit **G-SC/G-Over.** (7.80/7.17), surpassing BAGEL-14B in every column. Gains on extraction and removal support the source-preservation benefit of tokenizer-stage interaction (Section 5.3). The gap to Qwen-Image-20B is 0.20/0.39 on **G-SC/G-Over.**; Appendix Table 13 gives full results.

7 Conclusion

We presented HYDRA-X, a native UMM that unifies image and video tokenization within a single ViT. Tubelet causal attention, hierarchical temporal patchify, and dual-teacher distillation through a lightweight Decompressor equip the shared tokenizer with reconstruction fidelity and spatiotemporal semantics. Direct action-recognition and temporal-order probes verify that these semantics are retained in the compressed features consumed by the LLM. Reusing the temporal mechanism for source–target interaction improves editing consistency. Together, these designs support image/video understanding and generation, and image editing, within one shared framework.

Acknowledgments and Disclosure of Funding

This work is supported by the Basic Research Program of Jiangsu (No. BK20250009), the Fundamental Research Funds for the Central Universities (No. 020214380140), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118).

References

[1] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[2] Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025.

[3] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[4] Xuerui Qiu, Yutao Cui, Guozhen Zhang, Junzhe Li, JiaKui Hu, Xiao Zhang, Yang Li, Songtao Liu, Miles Yang, Yu Shi, et al. Hydra: Unifying multi-modal generation and understanding via representation-harmonized tokenization. *arXiv preprint arXiv:2603.15228*, 2026.

[5] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[6] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025.

[7] Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

[8] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025.

[9] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

[10] 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.

[11] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[12] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[13] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,

Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[14] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[15] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

[16] Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, KunPeng Du, Yi Wang, Limin Wang, et al. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. *arXiv preprint arXiv:2510.10575*, 2025.

[17] Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. *arXiv preprint arXiv:2512.13687*, 2025.

[18] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025.

[19] Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*, 2025.

[20] Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025.

[21] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.

[22] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024.

[23] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[24] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025.

[25] Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. *arXiv preprint arXiv:2503.19480*, 2025.

[26] Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. Autoregressive semantic visual reconstruction helps vlms understand better. *arXiv preprint arXiv:2506.09040*, 2025.

[27] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[28] Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*, 2025.

[29] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[30] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[31] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[32] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[33] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[34] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[35] JiaKui Hu, Shanshan Zhao, Qing-Guo Chen, Xuerui Qiu, Jialun Liu, Zhao Xu, Weihua Luo, Kaifu Zhang, and Yanye Lu. Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images. *arXiv preprint arXiv:2511.07222*, 2025.

[36] Junzhe Li, Sifan Zhou, Liya Guo, Xuerui Qiu, Linrui Xu, TingTing Long, Chun Fan, Ming Li, Hehe Fan, Jun Liu, et al. Unif2ace: A unified fine-grained face understanding and generation model. In *International Conference on Learning Representations*, volume 2026, pages 144143–144169, 2026.

[37] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025.

[38] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[39] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

[40] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025.

[41] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

[42] Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2024.

[43] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[44] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[45] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[46] Xiaoce Wang, Sifan Zhou, Kaifei Wang, Leli Xu, Xuerui Qiu, Tao He, and Ming Li. Why do DiT editors drift? plug-and-play low frequency alignment in VAE latent space. *arXiv preprint arXiv:2605.08250*, 2026.

[47] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

[48] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[49] Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Lang Lin, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. Internvideo-next: Towards general video foundation models without video-text supervision. *arXiv preprint arXiv:2512.01342*, 2025.

[50] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

[51] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[52] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[53] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[54] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[55] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[56] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209, 2021.

[57] Minesh Mathew, Viraj Bagal, Rubèn P Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706, 2022.

[58] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[59] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[60] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[61] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

[62] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[63] OpenAI. Gpt-4v(ision) system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[64] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[65] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[66] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024.

[67] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

[68] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[69] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[70] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[71] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[72] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[73] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[74] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.

[75] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[76] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[77] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[78] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[79] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[80] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[81] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[82] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24135–24146, 2025.

[83] Yicheng Xiao, Lin Song, Rui Yang, Cheng Cheng, Zunnan Xu, Zhaoyang Zhang, Yixiao Ge, Xiu Li, and Ying Shan. Haploomni: Unified single transformer for multimodal video understanding and generation. *arXiv preprint arXiv:2506.02975*, 2025.

[84] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[85] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[86] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

[87] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[88] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[89] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104, June 2025.

[90] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=EElFGvt39K.

[91] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[92] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014.

[93] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[94] JiaKui Hu, Tailai Chen, Yuqi Pan, Xuerui Qiu, Jialun Liu, Xiao Cao, Zhenxin Zhu, Guang Chen, Hangjun Ye, Bing Wang, and Yanye Lu. Learning to retrieve: Internalizing memory retrieval for video world models, 2026. URL https://arxiv.org/abs/2610.11444.

[95] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.

[96] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[97] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. *OpenAI technical report*, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

A Training Details

A.1 Tokenizer Training Loss

HYDRA-XTOK is designed as the visual interface of HYDRA-X: before any token reaches the LLM, it must be compact enough for generation, faithful enough for reconstruction, and semantic enough for understanding. We initialize Gen-ViT and Sem-ViT from SigLIP 2 [11]. The tokenizer is trained with a reconstruction term and a semantic distillation term:

$$\mathcal{L}_{\text{HYDRA-XTOK}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}},$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction term detailed below and $\mathcal{L}_{\text{dist}}$ aligns Sem-ViT features with the image and video teachers (Eq. 3).

To keep the compact latent both pixel-faithful and structurally stable, the reconstruction term $\mathcal{L}_{\text{rec}}$ encapsulates pixel-level recovery, perceptual fidelity, and latent space regularization. Specifically, it combines an L1 loss for direct pixel-space reconstruction, an LPIPS perceptual loss $\mathcal{L}_{\text{lpips}}$, an adversarial GAN loss $\mathcal{L}_{\text{gan}}$ to refine texture realism, and a Kullback–Leibler (KL) divergence penalty that aligns the posterior with a standard normal prior. The comprehensive reconstruction objective is formulated as:

$$\mathcal{L}_{\text{rec}} = \lambda_1\|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \lambda_{\text{perc}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{gan}}\mathcal{L}_{\text{gan}} - \lambda_{\text{KL}}\sum_{j=1}^{C}\left(1 + \boldsymbol{\rho}_j - \boldsymbol{\mu}_j^2 - \exp(\boldsymbol{\rho}_j)\right), \tag{7}$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are the original and reconstructed images, while $\boldsymbol{\mu}_j$ and $\boldsymbol{\rho}_j$ are the mean and log-variance of the compressed latent.

A.2 Tokenizer Pre-training

HYDRA-XTOK is trained in three progressive stages to balance foundational representation learning with high-fidelity generative quality:

Stage 1: Foundation Training. Initialized with SigLIP-2, HYDRA-XTOK first undergoes training on ImageNet-1.2M at 256×256 resolution. We then transition to mixed-resolution training, combining 256×256 videos with images ranging from 256 to 2048 pixels. This strategy empowers the tokenizer to generalize effectively to high-resolution video. We optimize the model for 300k iterations using AdamW with a peak learning rate of 2×10^{-4}, employing a hybrid SigLIP-2 / InternVideo teacher for distillation.

Stage 2: Decoder Refinement. To enhance texture realism and perceptual fidelity, we freeze the encoder and exclusively fine-tune the 27-layer ViT decoder. Adversarial training (GAN loss) is incorporated in this stage to significantly improve visual reconstruction.

Stage 3: Representation Harmonization. In the final stage, we first compute the channel-wise mean and standard deviation of the Gen-ViT latent features. We then freeze Gen-ViT and the decoder while unfreezing Sem-ViT. The Gen-ViT features are normalized before being fed into Sem-ViT and the decoder; during this process, only Sem-ViT is updated. This normalization eliminates feature heterogeneity between the two heads and establishes a unified, semantic-aware latent space capable of faithful reconstruction, which is crucial for downstream UMM tasks.

A.3 Native Unified Multimodal Models Pre-training

To cultivate the harmonized nature of HYDRA-X, we implement a three-stage progressive training strategy for the unified multimodal model. Detailed configurations and computational cost are summarised in Table 9.

Table 9: **Training details and computational cost of our HYDRA-X.** The HYDRA-XTOK pre-training takes an additional 24h on 256 GPUs. †Data Ratio denotes Text: Image Caption : Image Generation : Video Caption: Video SFT: Image SFT: Edit.

Setting	Stage 1	Stage 2	Stage 3
LR.	Vision Head: 10^{-4} Sem-ViT: 5×10^{-5}	Vision Head: 5×10^{-5} LLM & Sem-ViT: 2×10^{-5}	Vision Head: 5×10^{-5} LLM & Sem-ViT: 2×10^{-5}
Base Resolution	256	512	1024
Batch Size	1024	1024	1024
Tasks	Image (Und. & Gen.)	+ Video & Edit	+ Text Und.
Data Ratio†	0:1:3:0:0:0:0	0:2:6:1:1:0:0	1:0:3:3:0:1:3
Hardware	256 GPUs	512 GPUs	512 GPUs
Training Step	50K	200K	20K
Time Cost	∼10h	∼96h	∼24h

Stage 1: Unified Representation Alignment. To resolve the representation divergence at the input level, we freeze the LLM (Qwen2.5-7B-Instruct) and exclusively tune the vision components (projector, time-step embedding, and flow head). Utilizing 100M image–text pairs, this phase aligns the visual latent space with the linguistic domain, ensuring a coherent unified input representation.

Stage 2: Comprehensive Multimodal Pre-training. We unlock all parameters to facilitate harmonized co-promotion within a single unified stream. The model is jointly optimized on a balanced mix of 30M understanding samples and 30M generative samples (strategically filtered from Stage 1). We further incorporate approximately 2M image editing samples and 10M video samples into the joint training process. This full-parameter update ensures the compatibility of the learning process and allows the diverse tasks to mutually reinforce each other.

Stage 3: High-Quality Instruction Fine-tuning. The final stage focuses on high-fidelity refinement using curated datasets. For multimodal understanding (MMU), we employ 6M instruction-tuning samples sourced from LLaVA-OneVision [60] and Pixmo [89], alongside 1.2M video instruction-tuning samples from LLaVA-Video [90]. For generation, we utilize 10M aesthetic-filtered images (derived from Stage 2) and 6M high-fidelity synthetic images. We also train on high-quality image editing data to improve precise control.

A.4 Ablation Study Training Details

In our ablation studies, the evaluation covers three core capabilities with specific setups: (i) **Multimodal Understanding:** We train HYDRA-X using the LLaVA-1.5 multimodal understanding dataset [91] combined with the LLaVA-Video SFT dataset [90]. (ii) **Image Generation:** We use Qwen2.5-1.5B [13] as the base model, first training on 20M image-caption pairs, then further fine-tuning with the ImgEdit dataset for image editing capabilities. (iii) **Image Reconstruction:** We train on the ImageNet-1k (1.2M) dataset [92] for 150k iterations and assess quality using rFID.

B Visual Reconstruction

We benchmark HYDRA-XTOK under a unified protocol against three families of tokenizers: image-only generative VAEs, video VAEs, and joint image–video tokenizers. All inputs are resized and centre-cropped to 256×256 and metrics are computed with identical scripts. Table 10 reports PSNR, SSIM, and rFID/rFVD on ImageNet, DAVIS, and UCF.

To isolate the effect of architecture from that of training data, we additionally report HYDRA-XTOK †, a controlled variant trained strictly on ImageNet-1.2M to match the data budget of RAE† and VAVAE†. Under this matched-data setting, HYDRA-XTOK † outperforms RAE† and VAVAE† by a large margin on every ImageNet metric (e.g., $+5.26$ dB PSNR over VAVAE†). More strikingly, despite operating at *twice* the compression ratio of dedicated $8\times$ image VAEs and using strictly less training data, HYDRA-XTOK † still exceeds the strongest $8\times$ image VAE, FLUX.1, on ImageNet PSNR (32.96 vs. 32.86) and rFID (0.154 vs. 0.176), indicating that the holistic ViT design rather than the data scale drives the gain. The fully trained HYDRA-XTOK is the strongest unified tok-

Table 10: **Reconstruction comparison on ImageNet, DAVIS, and UCF.** All methods are evaluated with a unified protocol using their official implementations: inputs are resized and center-cropped to 256×256 and metrics are computed with identical scripts. Compression ratios are reported separately along the spatial (f_s) and temporal (f_t) axes; image-only tokenizers have f_t=1. Within each subgroup, **bold** marks the best result and underline marks the second-best. † indicates models trained strictly on the ImageNet-1.2M dataset.

Method	Compression		ImageNet			DAVIS			UCF		
	Spatial	Temporal	PSNR (↑)	SSIM (↑)	rFID (↓)	PSNR (↑)	SSIM (↑)	rFVD (↓)	PSNR (↑)	SSIM (↑)	rFVD (↓)
Generation-only Tokenizers											
SD-VAE [93]	8×	1×	26.26	0.745	0.606	–	–	–	–	–	–
RAE† [14]	16×	1×	18.05	0.500	2.040	–	–	–	–	–	–
FLUX.1 [dev] [78]	8×	1×	**32.86**	**0.917**	**0.176**	–	–	–	–	–	–
Qwen-Image [85]	8×	1×	32.18	0.899	1.459	–	–	–	–	–	–
VAVAE† [24]	16×	1×	27.70	0.798	0.279	–	–	–	–	–	–
Wan2.2 [10]	16×	4×	31.25	0.878	0.749	**27.64**	**0.820**	**14.78**	**36.11**	**0.961**	**4.15**
Unified Tokenizers											
OmniTokenizer [42]	8×	4×	26.74	0.824	1.023	24.30	0.737	113.56	29.20	0.931	38.15
Vila-U [62]	16×	1×	22.24	0.612	4.231	–	–	–	–	–	–
UniTok [9]	16×	1×	25.34	0.742	0.362	–	–	–	–	–	–
AToken-So/C (Stage 3) [28]	16×	4×	29.72	0.848	0.209	26.60	0.784	29.19	34.66	0.953	7.77
HYDRA-XTOK †	16×	1×	**32.96**	**0.905**	**0.154**	–	–	–	–	–	–
HYDRA-XTOK (Stage 3)	16×	4×	32.04	0.898	0.465	**28.19**	**0.835**	**11.61**	**36.88**	**0.967**	**3.11**

enizer on *every* video metric: it improves over the previous best AToken-So/C by +1.59 dB DAVIS PSNR and +2.22 dB UCF PSNR, while more than halving rFVD on both datasets (11.61 vs. 29.19 on DAVIS; 3.11 vs. 7.77 on UCF). On video benchmarks, HYDRA-XTOK also outperforms the dedicated 16× video VAE Wan 2.2 (+0.55 dB DAVIS PSNR, +0.77 dB UCF PSNR; rFVD reduced by 21% and 25% respectively), suggesting that a single holistic ViT with hierarchical patchify is a competitive alternative to cascaded image+video designs.

C Evaluation Details of Multi-modal Understanding Benchmarks

To comprehensively evaluate the perception and reasoning capabilities of HYDRA-X, we employ nine diverse benchmarks covering general understanding, expert knowledge, document/chart comprehension, and fine-grained visual perception. We benchmark on AI2D [50] (`test` split), MME [51] (`test` split), MMMU [52] (`val` split), OCRBench [53] (`test` split), MMBench [54] (`dev_en` split), RealWorldQA (`test` split), ChartQA [55] (`test` split), DocVQA [56] (`val` split), and InfoVQA [57] (`val` split). Table 5 compares HYDRA-X against open-source UMMs at a similar scale. Overall, HYDRA-X matches or exceeds 7B native UMM baselines on most reported metrics, including OCR- and chart-heavy tasks that simultaneously require fine-grained visual details (e.g., character strokes, table cells) and rich semantic structure (e.g., layout and relational reasoning), both of which HYDRA-XTOK's compact latent is designed to preserve.

D Tokenizer-Stage Source–Target Interaction: Visual Evidence

Section 5.3 argues that tokenizer-stage source–target interaction (STI)—routing the source $\mathbf{x}_c$ and target $\mathbf{x}_t$ jointly through a shared Sem-ViT with tubelet causal attention rather than encoding them independently—is the missing ingredient for identity-faithful image editing. Quantitatively, this single change recovers nearly 7 dB of source-reconstruction PSNR while leaving the rest of the architecture and parameter count untouched (Table 4). Figure 5 provides the qualitative counterpart.

We compare two variants of HYDRA-X that differ *only* in this routing step: **HYDRA-X-Indep** encodes the source and target through two independent Sem-ViT branches—the conventional pipeline shared by BAGEL, OmniGen2, and similar systems—while **HYDRA-X-STI** routes the pair through a single shared Sem-ViT with tubelet causal attention, treating $(\mathbf{x}_c, \mathbf{x}_t)$ as a length-2 clip. Both variants share the same Gen-ViT, the same LLM, and the same flow-matching head, with identical parameter count and training schedule.

The contrast is striking. In the still-life example (top row), HYDRA-X-Indep collapses into a fragmented mosaic where fruit positions, textures, and lighting are all hallucinated locally, whereas HYDRA-X-STI returns a near-pixel-perfect reproduction. The car example (bottom row) makes the failure mode of independent encoding explicit: the Indep variant "re-imagines" the car as a dif-

ferent vehicle, removing the driver and passenger and erasing the on-screen text; the STI variant preserves the entire scene including the people inside and the visible plate. These observations confirm the mechanism analysed in Section 5.3: in the conventional pipeline, the latent already loses identity-sensitive information *before* the LLM ever reads it, so even a perfectly reasoning LLM cannot recover the source faithfully. STI fixes this bottleneck inside the tokenizer at zero parameter cost, which is precisely what enables the consistent margin on identity-sensitive editing dimensions reported in Tables 8 and 13.

E Limitations

First, the current scale of training data and model parameters remains a bottleneck, potentially limiting the model's ability to capture the full complexity of high-dimensional video distributions. Second, resource constraints prevented us from exploring long video generation [94] and video editing, both of which are natural extensions of our holistic encoder. Finally, for a fair comparison, we instantiate HYDRA-X only on a 7B *dense* LLM; pairing our tokenizer with more advanced backbones, such as MoE [95] or MoT [96], represents a clear path to further amplify cross-task performance gains.

F Broader Impacts

As HYDRA-X introduces strong text-to-image generation capabilities within a unified framework, we acknowledge potential downstream risks. These include the generation of misleading or fabricated visual content (e.g., deepfakes), which could be exploited for disinformation or impersonation. To mitigate such risks, we advocate for the incorporation of content watermarking and provenance tracking mechanisms upon deployment, as well as adherence to responsible release practices such as gated model access and usage guidelines. We believe that advancing the scientific understanding of unified multimodal architectures carries substantial positive societal value, while the associated risks can be effectively managed through community-driven safety standards.

G Additional Main Results

This section provides the full per-category breakdown of the benchmarks summarised in Section 6, complementing the condensed tables in the main paper.

GenEval. Table 11 reports the per-category breakdown on GenEval [75], covering single-object, two-object, counting, color, position, and color-attribute prompts. The breakdown helps locate the compositional dimensions where each model is strongest.

Table 11: **Detailed image generation results on the GenEval benchmark [75].** Rows in gray indicate models with $\geq$ 14B parameters and are excluded from the ranking. † refers to methods using LLM rewriters.

Models	Size	Single Object	Two Objects	Count	Colors	Position	Color Attribute	Overall
Generation-only Models								
SD3-Med [45]	2B	**0.99**	**0.94**	0.72	0.89	0.33	0.60	0.74
FLUX.1 [Dev] [78]	12B	0.98	0.93	**0.75**	**0.93**	**0.68**	**0.65**	**0.82**
DALL-E 3 [97]	–	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Unified Multimodal Models								
TokenFlow-XL [18]	14B	0.95	0.60	0.41	0.81	0.16	0.24	0.55
SEED-X [81]	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
Ming-UniVision [61]	16B	1.00	0.93	0.59	0.93	0.92	0.70	0.85
BAGEL† [3]	14B	0.98	0.95	0.84	0.95	0.78	0.77	0.88
MetaQuery-XL [32]	7B	–	–	–	–	–	–	0.80
Blip3-o† [30]	8B	–	–	–	–	–	–	0.84
MUSE-VL [82]	7B	0.98	0.64	0.54	0.72	0.25	0.31	0.57
Janus-Pro [38]	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Show-o2† [1]	7B	**1.00**	0.87	0.58	**0.92**	0.52	0.62	0.76
HYDRA† [4]	7B	**1.00**	**0.97**	0.68	0.91	0.81	**0.80**	0.86
HYDRA-X	7B	0.99	0.95	**0.83**	0.90	**0.84**	0.75	**0.88**

WISE. Table 12 reports the per-category breakdown on WISE [76], which probes world knowledge across culture, time, space, biology, physics, and chemistry, and is therefore complementary to the geometric and compositional probes of GenEval.

Table 12: **Detailed image generation results on the WISE benchmark [76].** Rows in gray indicate models with $\geq$ 14B parameters and are excluded from the ranking.

Models	Size	Culture	Time	Space	Biology	Physics	Chemistry	Overall
Generation-only Models								
FLUX.1 [Dev] [78]	12B	**0.48**	**0.58**	**0.62**	0.42	0.51	**0.35**	**0.50**
SD3.5-Large [45]	8B	0.44	0.50	0.58	**0.44**	**0.52**	0.31	0.46
Unified Multimodal Models								
BAGEL [3]	14B	0.44	0.55	0.68	0.44	0.60	0.39	0.52
VILA-U-7B [62]	7B	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Janus-Pro-7B [38]	7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3-Gen-8B [84]	8B	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Show-o2 [1]	7B	0.40	0.45	0.58	0.39	0.53	0.34	0.44
HYDRA [4]	7B	0.52	0.53	0.68	0.47	0.58	0.38	0.53
HYDRA-X	7B	**0.53**	**0.57**	**0.72**	**0.53**	**0.64**	**0.40**	**0.56**

ImgEdit-Bench. Table 13 provides the full per-dimension breakdown on ImgEdit-Bench [47], spanning nine instruction-guided editing operations from object addition and removal to background replacement, style transfer, and compositional edits.

VBench. Table 14 expands the QS/SS/Total summary in the main paper to all fourteen VBench [77] dimensions, separately probing visual quality, motion smoothness, dynamic degree, semantic correctness, and compositional reasoning.

Table 13: **Detailed image editing results on the ImgEdit-Bench [47].** Editing dimensions: **Add**, **Adj.** (Alter), **Ext.** (Extract), **Rep.** (Replace), **Rm.** (Remove), **Bg.** (Background), **Sty.** (Style), **Hyb.** (Compose), **Act.** (Action). Rows in gray indicate models with $\geq$ 14B parameters and are excluded from the ranking.

Models	Size	Add	Adj.	Ext.	Rep.	Rm.	Bg.	Sty.	Hyb.	Act.	Overall
Generation-only Models											
FLUX.1 Kontext [Pro] [78]	12B	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
Qwen-Image [85]	20B	**4.38**	**4.16**	**3.43**	**4.66**	**4.14**	**4.38**	**4.81**	**3.82**	**4.69**	**4.27**
Unified Multimodal Models											
BAGEL [3]	14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
OmniGen [86]	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
UniWorld-V1 [31]	12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [87]	4B	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	**4.68**	3.44
HYDRA-X	7B	**4.49**	**4.27**	**4.04**	**4.41**	**4.38**	**4.30**	4.77	**3.43**	4.32	**4.34**

Table 14: **Detailed video generation results on VBench [77].** Column abbreviations: **QS**: Quality Score, **SS**: Semantic Score, **SC**: Subject Consistency, **BC**: Background Consistency, **MS**: Motion Smoothness, **DD**: Dynamic Degree, **AQ**: Aesthetic Quality, **IQ**: Imaging Quality, **OC**: Object Class, **MO**: Multiple Objects, **HA**: Human Action, **C**: Color, **SR**: Spatial Relationship, **S**: Scene.

Models	Size	QS	SS	SC	BC	MS	DD	AQ	IQ	OC	MO	HA	C	SR	S	Total
Generation-only Models																
CogVideoX [79]	5B	82.75	**77.04**	96.23	96.52	96.92	70.97	**61.98**	62.90	**85.23**	62.11	**99.40**	82.81	66.35	53.20	81.61
Hunyuan Video [80]	13B	**85.07**	76.88	**97.22**	**97.60**	**99.05**	**71.94**	60.28	**67.24**	83.48	**66.71**	94.40	**89.79**	**72.13**	**54.46**	**83.43**
Unified Multimodal Models																
VILA-U [62]	7B	76.26	65.04	–	–	–	–	–	–	–	–	–	–	–	–	74.01
HaploOmni [83]	7B	–	–	96.40	97.60	96.80	65.30	–	–	–	–	–	–	–	–	78.10
Emu3 [84]	8B	–	–	95.32	**97.69**	**98.93**	**79.27**	59.64	–	86.17	44.64	77.71	–	68.73	37.11	80.96
Show-o2 [1]	1.5B	82.10	78.31	**97.28**	96.78	98.25	40.83	**65.15**	**67.06**	94.81	76.01	95.20	80.89	62.61	57.67	81.34
HYDRA-X	7B	**83.97**	**81.57**	95.70	95.88	92.69	35.42	63.73	65.23	**96.52**	**76.68**	**99.00**	**87.37**	**73.88**	**69.74**	**83.49**

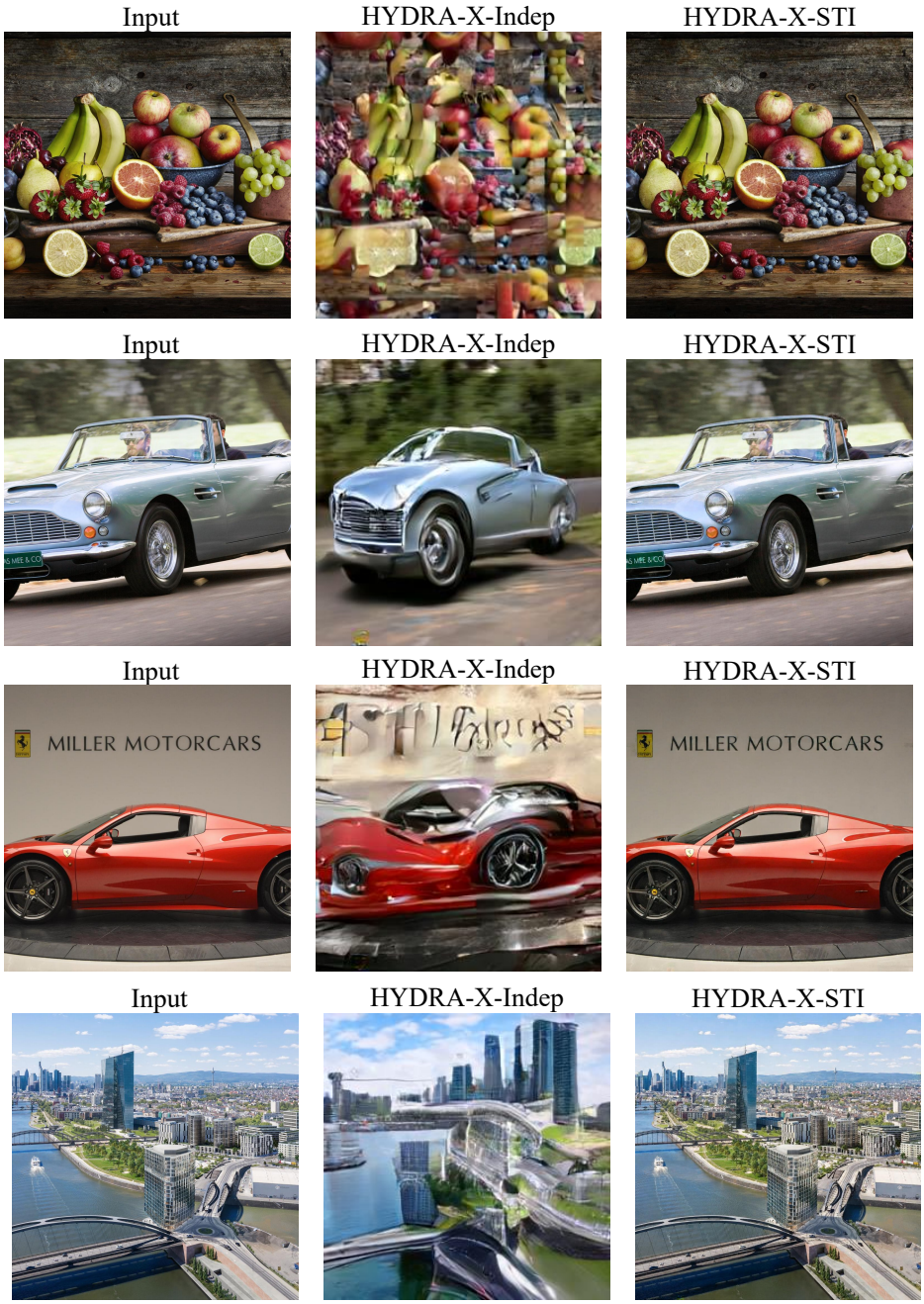


Figure 5: **Qualitative effect of tokenizer-stage source–target interaction.** Source-image reconstruction produced by HYDRA-X-Indep (independent Sem-ViT encoding of source and target, the conventional pipeline) versus HYDRA-X-STI (joint encoding through tubelet causal attention, our proposal). The two variants share every other architectural component. HYDRA-X-STI preserves identity-sensitive details (object layout, characters, on-screen text) that HYDRA-X-Indep loses, despite both pipelines using the same LLM and the same number of parameters.

H Qualitative Comparisons

This section presents qualitative results across the five tasks supported by HYDRA-X. We compare against representative baselines drawn from both unified multimodal models and task-specialised systems, and organise the comparisons by task and resolution.

H.1 Image Reconstruction at 512×512

We first inspect reconstruction fidelity at the standard 512×512 resolution. The comparison spans three families of baselines: dedicated image VAEs (FLUX), unified tokenizers built into UMMs (MingTok, AToken), and the recently proposed RAE. The visual difference makes texture, fine-edge, and small-text fidelity directly comparable.

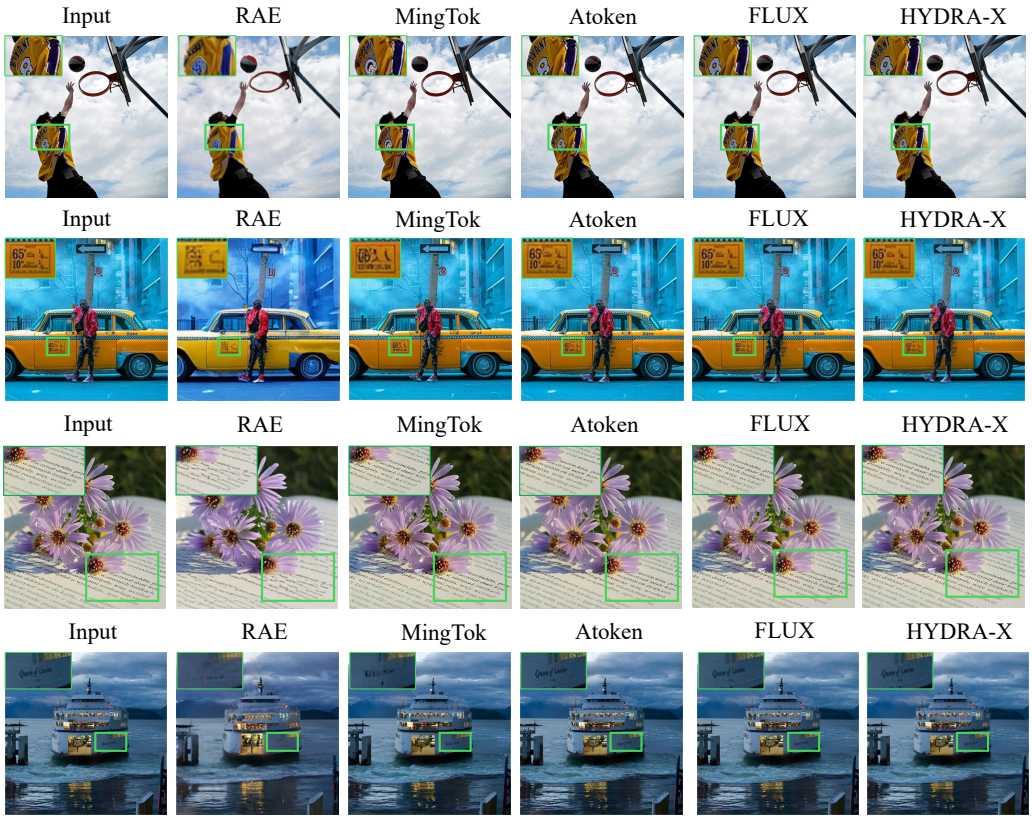


Figure 6: **Qualitative reconstruction comparison at 512×512.** We compare HYDRA-X against RAE [14], MingTok [61], AToken [28], and FLUX [78].

H.2 Image Reconstruction at 1280×768

To stress-test generalisation beyond the training resolution, we additionally compare reconstructions at a high resolution of 1280×768 and include the dedicated video VAE Wan 2.2 alongside the image-only baselines. This setting exposes how each tokenizer handles dense fine details such as text, foliage, and small structural elements when the spatial token budget is stretched.

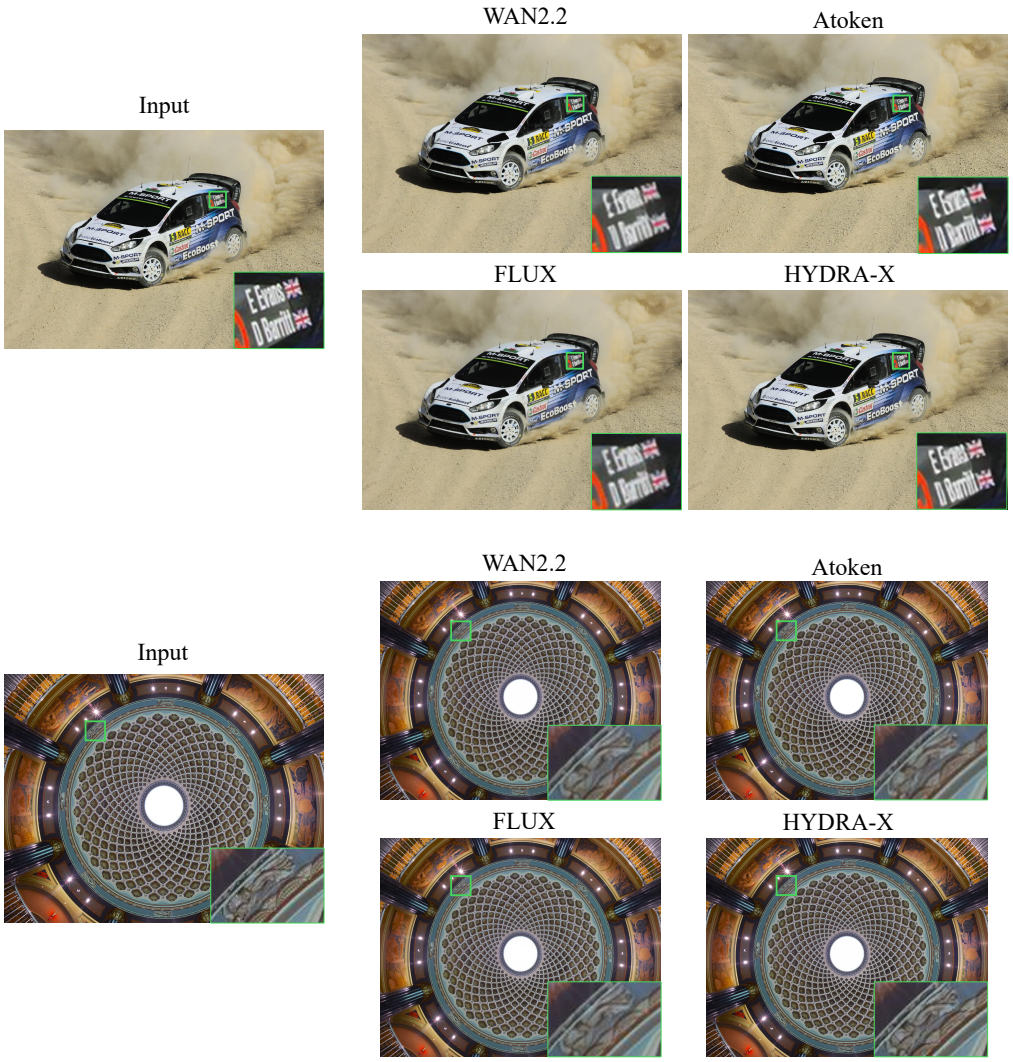


Figure 7: **Qualitative reconstruction comparison at** 1280×768**.** We compare HYDRA-X against Wan 2.2 [10], AToken [28], and FLUX [78].

H.3 Video Reconstruction at 512×512

Beyond static images, we visualise temporally consecutive frames reconstructed by HYDRA-X against the dedicated video VAE Wan 2.2 and the joint image–video tokenizer AToken. This helps assess whether HYDRA-XTOK's holistic ViT preserves motion-sensitive cues such as object boundaries and inter-frame consistency.

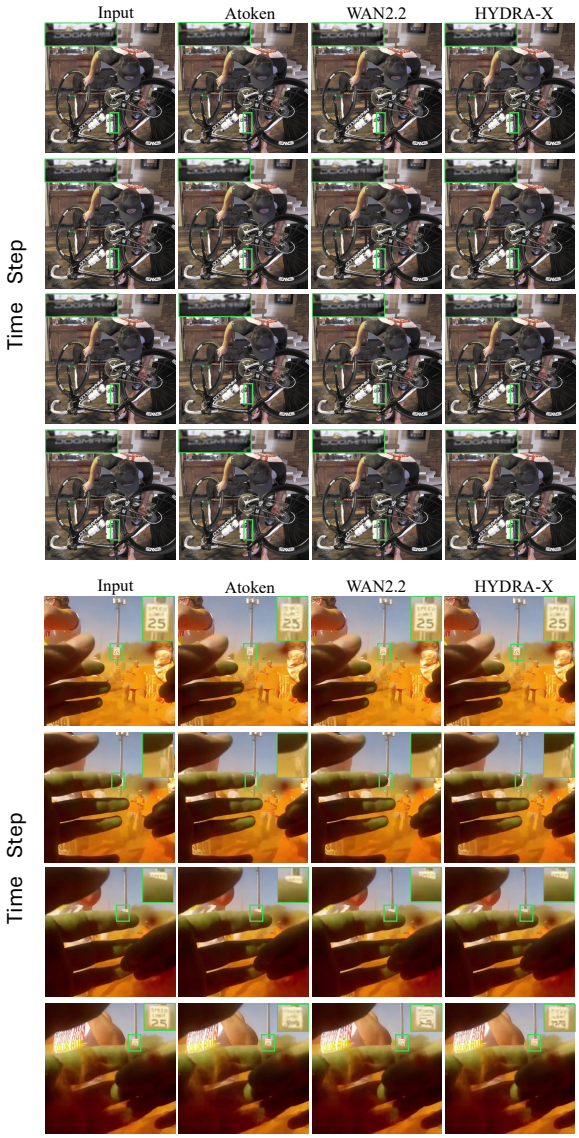


Figure 8: **Qualitative video reconstruction comparison.** We compare HYDRA-X against Wan 2.2 [10] and AToken [28].

H.4 Image Generation

We provide qualitative text-to-image samples produced by HYDRA-X spanning a diverse range of prompts—from realistic photography and stylised illustration to compositional and knowledge-driven scenes—to characterise the model's coverage and aesthetic quality.



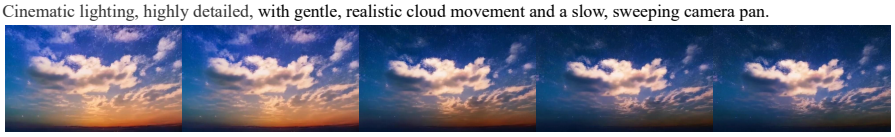
Figure 9: **Qualitative image generation results from HYDRA-X.**

H.5 Video Generation

We present text-to-video samples across varied subjects, scenes, and motions, illustrating temporally coherent synthesis from the holistic latent with the shared UMM backbone.



A majestic, photorealistic wide shot of snow-capped mountains under a bright, clear blue sky. Fluffy white clouds elegantly form the word 'HYDRA-X' in the center of the sky. The text clouds blend naturally with scattered, regular clouds around them. Cinematic lighting, highly detailed, with gentle, realistic cloud movement and a slow, sweeping camera pan.



A grand, expansive vista of a dramatic sunset blending into a starry night sky. The top is a clear blue with the Milky Way and stars visible, while the bottom horizon glows with rich orange or red hues. Fluffy clouds are scattered across the scene. Timelapse-style camera showing the gradual shift of the sky and stars. Gentle camera pan.



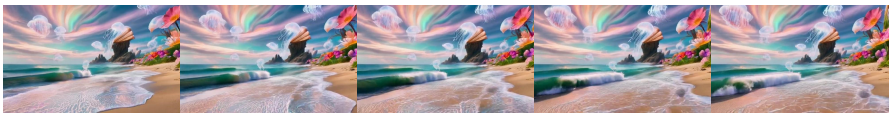
A wide-angle, cinematic shot of a vast golden desert with detailed, wind-swept sand dunes in the foreground. Above, a breathtakingly clear, deep blue night sky filled with bright stars and a vividly glowing Milky Way galaxy. In the distant horizon, a mysterious, massive dark silhouette resembling a giant hand or abstract rock formation rises towards the stars. Surreal atmosphere, highly detailed, with a smooth, slow camera pan across the silent landscape.



A close-up view of a brawny humanoid werewolf-like creature with grey fur, standing in a blurry forest setting. The creature looks directly into the lens with an intense, weary expression. It holds its hands firmly around an ornate, branch-decorated arrow with a red feather that is shot into its bloodied chest. Cinematic, dramatic lighting. Subtle chest movement as it breathes.



A smiling female astronaut floating in zero gravity inside a space station, her blonde hair drifting freely in the air, with a view of Earth and space through the window behind her.



A surreal dreamscape of a peaceful beach with glowing, translucent jellyfish floating in a pastel aurora sky, surrounded by giant colorful flowers and gentle ocean waves.



A beautiful watercolor-style animation of a great white shark swimming gracefully underwater above a vibrant coral reef.



A magical fantasy landscape featuring a glowing cyan river, vibrant colorful flora, a majestic golden tree, and glowing mountains under a purple starry sky with two moons.

Figure 10: **Qualitative video generation results from HYDRA-X.**

H.6 Image Editing

Finally, we compare HYDRA-X against representative editing systems on a set of instruction-guided edits. The baselines include both unified multimodal models (BAGEL, OmniGen2) and editing-specialised generators (Qwen-Image-Edit, Step1X-Edit), allowing readers to gauge identity preservation, instruction adherence, and visual quality side-by-side.

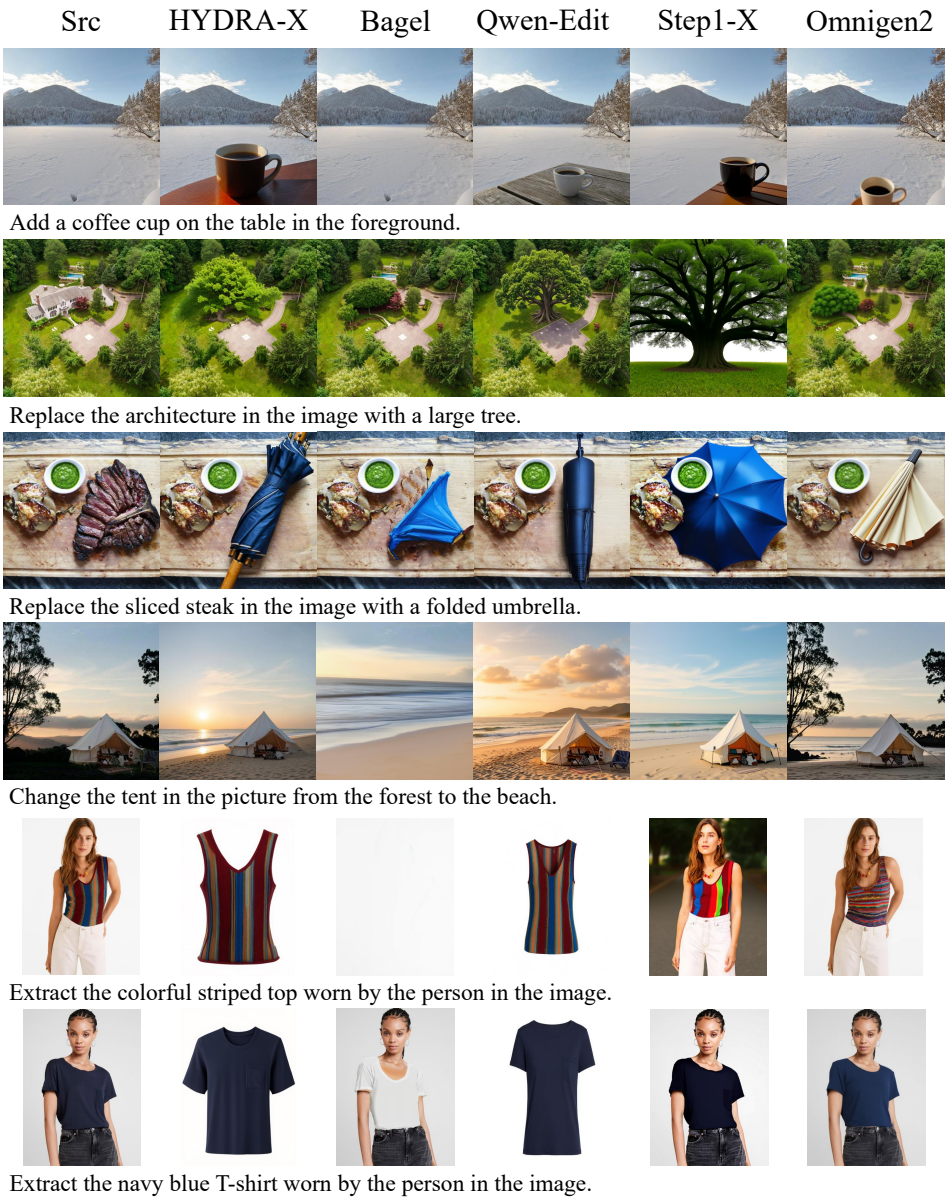


Figure 11: **Qualitative editing comparison.** We compare HYDRA-X against BAGEL [3], Qwen-Image-Edit [85], Step1X-Edit [88], and OmniGen2 [87].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions — (i) the holistic visual tokenizer HYDRA-XTOK for joint image-and-video tokenization in a single ViT, (ii) the native UMM HYDRA-X unifying five visual tasks, and (iii) tokenizer-stage source-target interaction for image editing — each empirically substantiated in Sections 4–6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E discusses algorithmic and empirical limitations, including the short temporal receptive field induced by the 2-frame tubelet design, the current restriction of source-target interaction to paired image editing, the absence of long-form video generation and high-resolution video editing experiments, and the use of a 7B dense LLM backbone under available compute.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical study of unified multimodal tokenization and modelling; it does not contain formal theorems, lemmas, or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the main architecture, ablation protocols, evaluation benchmarks, model scales, LLM backbone (Qwen2.5-7B-Instruct), distillation teachers (SigLIP 2-SO400M and InternVideo-Next-L), bottleneck dimension, and the three-stage tokenizer and UMM training recipes in Sections 4–6 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code, training scripts, data-processing scripts, and pretrained checkpoints are not released at submission time. We plan to open-source the tokenizer code, training scripts, model weights, and model card after the anonymity period, subject to license and safety review. The current paper provides architectural, training, and evaluation details, but not open access to the full artifacts needed for faithful reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation details (model scales, LLM backbone, tokenizer architecture, Decompressor, distillation teachers, and bottleneck dimension) are stated in Section 6. Appendix A provides the tokenizer pre-training recipe and the three-stage UMM training schedule, including data scales, task mixtures, resolutions, batch sizes, optimizer information, learning rates, training steps, and wall-clock costs. Evaluation benchmarks and public splits are described in Section 6 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report single-run scores from the official evaluation suites for GenEval, WISE, VBench, MVBench, Video-MME, ImgEdit-Bench, GEdit-Bench, and related benchmarks. Due to the computational cost of training 7B-scale UMMs, we do not report multiple training seeds, confidence intervals, or formal significance tests. This is a limitation of the empirical evidence.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A reports the number of GPUs, training steps, batch sizes, base resolutions, and wall-clock time for the three-stage HYDRA-X training pipeline, and Table 1 reports tokenizer latency for $17\times512\times512$ video clips.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms with the NeurIPS Code of Ethics. We rely on publicly available pretrained models (SigLIP 2, InternVideo-Next, and Qwen2.5) and public benchmark datasets, do not collect new human-subject data, and discuss potential misuse in Appendix F.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F discusses positive impacts (advancing scientific understanding of unified multimodal architectures) and negative impacts (deepfake or disinformation risks from improved image/video generation), and outlines mitigations such as watermarking, provenance tracking, gated release, and usage guidelines.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Appendix F describes planned safeguards for releasing the trained HYDRA-X checkpoint, including gated access, usage guidelines, content watermarking, and provenance tracking. We will additionally follow the release terms and safety policies of the pretrained models used to build the system.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The pretrained models and tokenizers used by the paper, including SigLIP 2 [11], InternVideo-Next [49], Qwen2.5-Instruct [13], and AToken [28], are cited. Public benchmark datasets used for evaluation, including GenEval, WISE, VBench, MVBench, Video-MME, LongVideoBench, LVBench, AI2D, MME, MMMU, OCR-Bench, MMBench, ChartQA, DocVQA, InfoVQA, ImgEdit-Bench, GEdit-Bench, ImageNet, DAVIS, and UCF, are cited in the corresponding experimental sections. We use these assets for research evaluation and follow their public licenses and terms of use; no third-party dataset is redistributed with this submission.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper proposes a new tokenizer architecture and a new model, but does not release new checkpoints, datasets, code packages, or other standalone assets at submission time. We plan to provide a model card, training code, and weights after the anonymity period, subject to license and safety review.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects; all evaluations are performed on public benchmarks with automated metrics.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects, so no IRB approval is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use the publicly available Qwen2.5-7B-Instruct [13] as the autoregressive backbone of HYDRA-X, declared in Section 6 and Appendix A. The LLM is used as a standard pretrained backbone shared by recent native UMMs and is not employed for writing, editing, or formatting of the manuscript itself.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.