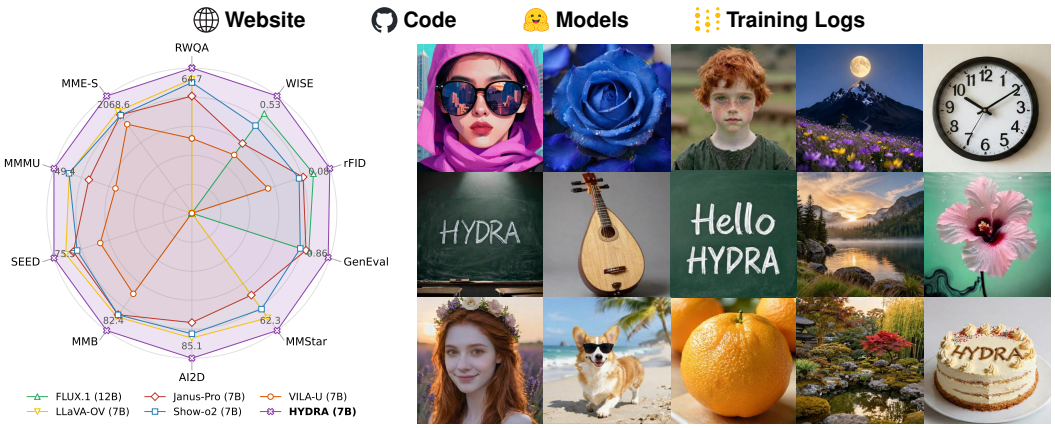
HYDRA: Representation Harmonized Tokenization for Multimodal Generation and Understanding

Xuerui Qiu[1,2,3*] **Yutao Cui**[2*] **Guozhen Zhang**[2,4*] **Junzhe Li**[2,5] **Yaqi Zhao**[2,5] **Xiao Zhang**[2]
Yang Li[2] **Songtao Liu**[2] **Miles Yang**[2] **Yu Shi**[3†] **Zhao Zhong**[2†] **Liefeng Bo**[2]

[1]Institute of Automation, Chinese Academy of Sciences [2]Tencent Hunyuan
[3]Zhongguancun Academy [4]Nanjing University [5]Peking University



(a) Comparison with SOTA. (b) Image generation results.

Figure 1: **Multimodal understanding and image generation results.** (a) HYDRA outperforms unified and task-specific models across diverse benchmarks. (b) HYDRA generates high-fidelity images with strong semantic alignment.

Abstract

Unified Multimodal Models struggle to bridge the fundamental gap between the abstract representations needed for visual understanding and the detailed primitives required for generation. Existing approaches typically compromise by employing decoupled encoders, stacking representation encoder atop VAEs, or utilizing discrete quantization. However, these methods often disrupt information coherence and lead to optimization conflicts. To this end, we introduce HYDRA-TOK, a representation-harmonized pure ViT in the insight that visual modeling should evolve from generation to understanding. HYDRA-TOK reformulates the standard backbone into a progressive learner that transitions from a Gen-ViT, which captures structure-preserving primitives, to a Sem-ViT for semantic encoding. Crucially, this transition is mediated by a Generation-Semantic Bottleneck (GSB), which compresses features into a low-dimensional space to filter noise for robust synthesis, then restores dimensionality to empower complex semantic comprehension. Built upon this foundation, we present HYDRA, a native unified framework integrating perception and generation within a single parameter space. Extensive experiments establish HYDRA as a new state-of-the-art. It sets a benchmark in visual reconstruction (rFID 0.08) and achieves top-tier generation performance on GenEval (0.86), DPG-Bench (86.4), and WISE (0.53), while simultaneously outperforming previous native UMMs by an average of $\sim$10.0 points across eight challenging understanding benchmarks.

*These authors contributed equally to this work.
†Corresponding authors: Yu Shi (shiyu@bza.edu.cn) and Zhao Zhong (caesarqin@tencent.com).

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

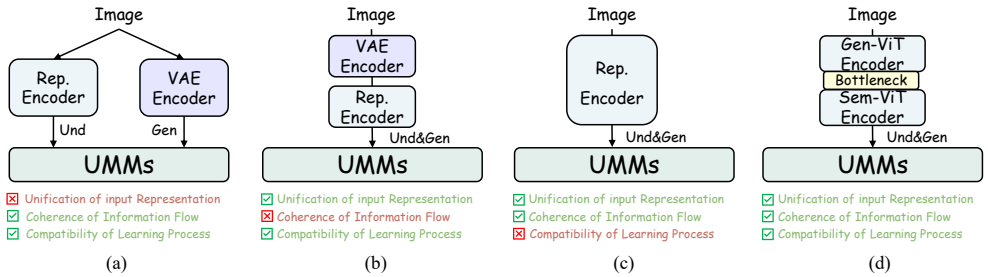


Figure 2: **Representation schemes in native unified multimodal models.** (a) Decoupled Encoder [1, 2]: Uses separate encoders (a VAE and a representation encoder) for generation and understanding. (b) Sequential Encoder [3]: Feeds the outputs of the VAE directly into the representation encoder in a cascaded manner. (c) Single Encoder [4, 5]: Unifies both tasks within a standalone representation encoder. (d) Our Proposed Representation-Harmonized ViT Design: Employs a single ViT backbone with a bottleneck module to harmonize feature learning across both tasks.

1 Introduction

Unifying visual understanding and generation has emerged as a pivotal frontier in multimodal intelligence [1, 6, 7]. Native unified multimodal models (UMMs) [5, 8, 3] advocate for direct decoding within a unified parameter space, demonstrating superior synergy over composite UMMs [9–11]. However, achieving rational unification is hindered by a fundamental representational divergence, stemming from the fact that understanding and generation constitute inverse tasks with conflicting demands: the former necessitates high-level semantic abstractions, whereas the latter requires compact structural primitives for fine-grained synthesis [12, 13]. This intrinsic conflict forces existing frameworks into disjointed, asymmetric designs, significantly increasing architectural complexity and optimization difficulty.

We attribute this dilemma primarily to the structural limitations of existing image tokenizers, which fail to simultaneously satisfy the three critical criteria illustrated in Fig. 2. First, decoupled paradigms that employ separate encoders for understanding and generation [1, 6] inherently lack *Unification of Input Representation*, relying on disjoint features that sever the synergy between the two tasks. Second, sequential architectures that stack representation encoders atop VAEs [3, 14] theoretically unify the input but compromise the *Coherence of Information Flow* due to the significant representation mismatch between the generative VAE latent space and the semantic features required by the representation encoder. Third, while utilizing a single shared representation encoder [4, 5, 15] attempts to solve this, it often suffers from poor *Compatibility of Learning Process*, where the conflicting objectives of high-frequency detail preservation and semantic abstraction lead to optimization difficulties. Consequently, current methods face an unavoidable trade-off: they either sacrifice generative fidelity, lose semantic alignment, or struggle to converge on a shared representation.

To address this, we propose HYDRA-TOK, a representation-harmonized pure ViT framework. Distinct from previous methods that employ convolutional VAEs for visual tokenization, HYDRA-TOK jointly trains reconstruction and understanding capabilities natively within a single ViT backbone. Our core design principle is built on a key insight: a compact feature space capable of reconstructing input data can serve as a robust foundation for semantic understanding. This reconstruction task functions as an information bottleneck, compelling the compact feature to discard extraneous details and instead acquire a vocabulary of dense, structural primitives. These primitives provide a solid basis, thereby enabling the model to construct semantic abstractions from the ground up.

Driven by this insight, we reformulate a ViT-based representation encoder [16, 17] into a progressive learner that transitions from a Generation vision transformer (Gen-ViT), which captures structure-preserving primitives for high-fidelity synthesis, to a Semantic vision transformer (Sem-ViT). To unify these distinct objectives within a single model, we introduce the Generation-Semantic Bottleneck (GSB). The GSB is architected around a novel compress-and-reconstruct operation, creating an information bottleneck that fosters both high-level semantic abstraction and detailed generative fidelity. Specifically, GSB first compresses features into a compact low-dimension space to filter out noise component, then reconstructs them to the original space for subsequent semantic distillation. In this manner, the compact features can both encode structured details while maintaining semantic awareness.

Built upon HYDRA-TOK, we present HYDRA, a unified framework that achieves complete architectural and representational unification. Leveraging the coherent visual representations provided by HYDRA-TOK, visual signals are processed as sequences via a dual-head mechanism, employing an autoregressive head for text prediction and a rectified flow matching head for image generation. Extensive experiments confirm that HYDRA achieves superior performance, highlighting the harmony between understanding and generation during joint training. In terms of multimodal understanding, HYDRA outperforms existing native UMMs by an average margin of approximately 10.0 points across eight benchmarks. Meanwhile, it establishes a new benchmark in visual reconstruction with a remarkable rFID of 0.08, providing a robust foundation that facilitates state-of-the-art generation records: 0.86 on GenEval [18], 86.4 on DPG-Bench [19], and 0.53 on WISE [20]. Additionally, we find that joint training consistently outperforms separate training for both generation and understanding, validating the effectiveness of our harmonized representation. Our contributions are summarized as follows:

- We propose HYDRA-TOK, a representation-harmonized pure ViT that resolves the understanding-generation conflict via a progressive learner, unifying input representations without quantization errors.
- We present HYDRA, a native unified framework that integrates understanding and generation within a single parameter space, utilizing a dual-head mechanism for seamless task execution.
- Empirical results demonstrate that HYDRA outperforms native UMMs by $\sim$10.0 points on understanding benchmarks and achieves state-of-the-art performance on GenEval, DPG-Bench, and WISE.

2 Method

We present HYDRA, a unified framework harmonizing visual understanding and generation. Central to our approach is HYDRA-TOK, a pure ViT grounded in a key insight: a compact feature space capable of reconstructing inputs serves as a robust foundation for semantic understanding. Adopting a functionally progressive learner design (Fig. 3), we transition continuously from structure-preserving primitives (Gen-ViT) to semantic abstractions (Sem-ViT). This is realized via the Generation-Semantic Bottleneck (GSB) and its novel compress-and-reconstruct operation. By compressing features to filter noise and reconstructing them for distillation, the GSB effectively balances generative fidelity with semantic awareness.

2.1 HYDRA-TOK

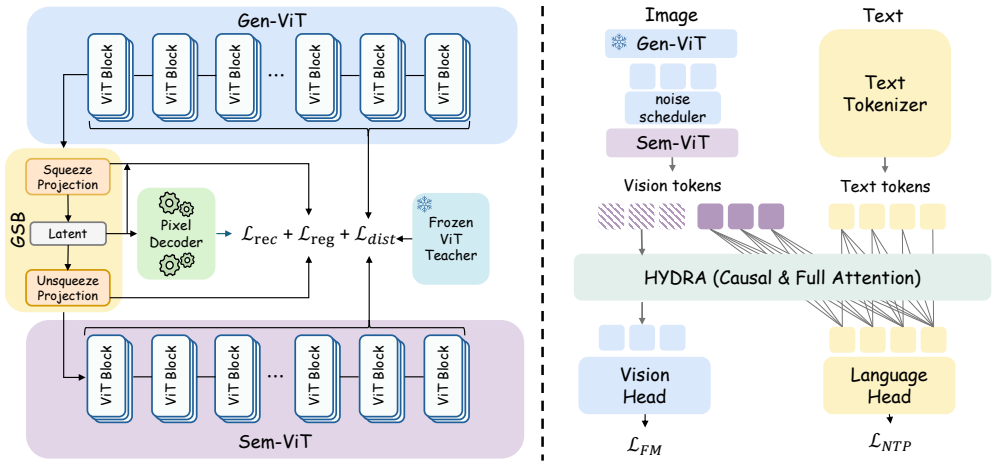
Traditional tokenizers face a rigid trade-off between preserving semantic depth and maintaining structural detail. To resolve this issue, HYDRA-TOK reformulates the complete vision transformer backbone into three functionally distinct yet continuous components, effectively followed by a lightweight flow-based decoder.

Generation Vision Transformer (Gen-ViT). Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, we first flatten and project non-overlapping patches into continuous embeddings $\mathbf{H}_0 \in \mathbb{R}^{N\times D}$, where N and D correspond to the number of tokens and dimension. The initial stage, Gen-ViT, is tasked with extracting low-level structural primitives essential for generation. Unlike standard encoders that aggressively compress spatial information, Gen-ViT preserves fine-grained spatial covariance:

$$\mathbf{H}_{\texttt{mid}} = \text{Gen-ViT}(\mathbf{H}_0) = \Phi_{L_{\texttt{gen}}} \circ \cdots \circ \Phi_1(\mathbf{H}_0), \tag{1}$$

where Φ_l denotes the l-th transformer block. This phase ensures that the latent space retains the structural foundation required for high-fidelity synthesis.

Generation-Semantic Bottleneck (GSB). To transition from structure-preserving primitives to semantic abstractions, we introduce the GSB block. Built on a compress-and-reconstruct operation, GSB serves as an information bottleneck that filters extraneous noise to balance the conflicting demands of understanding and generation. As evidenced by our ablation (Fig. 4), while higher dimensions (C) aid reconstruction and understanding, they cause generation performance to collapse (e.g., at $C \geq 256$). This confirms that excessive dimensionality introduces redundancy that disrupts generative stability.



(a) HYDRA-Tok (b) HYDRA

Figure 3: **Training process illustration for HYDRA-TOK and HYDRA.** (a) HYDRA-TOK functions as a progressive learner, bridging the gap between reconstruction and understanding. It employs a Generation-Semantic Bottleneck (GSB) to execute a unique compress-and-reconstruct operation, effectively filtering noise to transition from structure-preserving primitives (Gen-ViT) to semantic abstractions (Sem-ViT). (b) HYDRA unifies these representations using a dual-head architecture, effectively combining autoregressive text prediction and rectified flow matching.

To resolve this, GSB acts as a stabilization pivot by first compressing the intermediate features $\mathbf{H}_{\texttt{mid}}$ into a compact probabilistic space via a lightweight projector $\mathbf{W}_{\texttt{proj}} \in \mathbb{R}^{D\times 2C}$, where $C \ll D$ (typically 64):

$$[\boldsymbol{\mu}, \boldsymbol{\rho}] = \mathbf{H}_{\texttt{mid}}\mathbf{W}_{\texttt{proj}}, \quad \mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \exp(0.5\boldsymbol{\rho}), \tag{2}$$

where $\boldsymbol{\mu}, \boldsymbol{\rho} \in \mathbb{R}^{N\times C}$ represent the mean and log-variance, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is reparameterization noise.

To structure this latent space, we impose a KL divergence loss that aligns the posterior with a standard normal prior:

$$\mathcal{L}_{\texttt{KL}} = -\frac{1}{2}\sum_{j=1}^{C}\left(1 + \boldsymbol{\rho}_j - \boldsymbol{\mu}_j^2 - \exp(\boldsymbol{\rho}_j)\right). \tag{3}$$

To maintain a coherent flow of information under compression and provide a sufficient foundation for subsequent semantic extraction, we introduce a consistency loss $\mathcal{L}_{\texttt{cos}}$. This loss forces the unprojected features $\mathbf{H}_{\texttt{bn}} = \boldsymbol{\mu}\mathbf{W}^{\texttt{und}}_{\texttt{unproj}}$ to maintain directional alignment with the pre-bottleneck features $\mathbf{H}_{\texttt{mid}}$:

$$\mathcal{L}_{\texttt{cos}} = 1 - \frac{\mathbf{H}_{\texttt{mid}} \cdot \mathbf{H}_{\texttt{bn}}}{\|\mathbf{H}_{\texttt{mid}}\|_2 \|\mathbf{H}_{\texttt{bn}}\|_2}. \tag{4}$$

Based on our experiments, we set the regularization weights to $\lambda_{\texttt{KL}} = 10^{-4}$ and $\lambda_{\texttt{cos}} = 1.0$. The composite bottleneck objective is defined as $\mathcal{L}_{\texttt{reg}} = \lambda_{\texttt{KL}}\mathcal{L}_{\texttt{KL}} + \lambda_{\texttt{cos}}\mathcal{L}_{\texttt{cos}}$.

Semantic Vision Transformer (Sem-ViT). As the final stage of our functionally progressive learner, Sem-ViT acts as a deep non-linear mapper. Its primary role is to transition the structural foundations (encoded in $\mathbf{H}_{\texttt{bn}}$) into a high-dimensional semantic space, thereby achieving rational disentanglement:

$$\mathbf{H}_{\texttt{out}} = \text{Sem-ViT}(\mathbf{H}_{\texttt{bn}}) = \Phi_{L_{\texttt{total}}} \circ \cdots \circ \Phi_{L_{\texttt{gen}}+1}(\mathbf{H}_{\texttt{bn}}). \tag{5}$$

To ensure robust representation learning across the entire hierarchy, we employ semantic self-distillation on both Gen-ViT and Sem-ViT. We align the intermediate features from distinct depths of the student with a frozen, pre-trained ViT [16] via cosine similarity maximization:

$$\mathcal{L}_{\texttt{dist}} = \sum_{l\in\mathcal{S}_{\texttt{gen}}\cup\mathcal{S}_{\texttt{sem}}}\left(1 - \frac{\mathbf{H}^{(l)}(\mathbf{x}) \cdot \mathcal{T}^{(l)}(\mathbf{x})}{\|\mathbf{H}^{(l)}(\mathbf{x})\|_2 \|\mathcal{T}^{(l)}(\mathbf{x})\|_2}\right), \tag{6}$$

where $\mathcal{S}_{\texttt{gen}}$ and $\mathcal{S}_{\texttt{sem}}$ denote the selected layers from Gen-ViT and Sem-ViT, respectively.

Pixel Flow Decoder. To unburden the backbone, we employ a lightweight decoder $\mathbf{v}_\theta$ that utilizes flow matching to recover high-frequency details. Conditioned on the sampled latent $\mathbf{c} = \mathbf{z}$ from the GSB (Eq. 2), it learns to regress the velocity field by minimizing:

$$\mathcal{L}_{\texttt{FM}} = \mathbb{E}_{t,\mathbf{x},\boldsymbol{\epsilon}} \left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{x})\|^2 \right]. \tag{7}$$

To further enhance perceptual fidelity, we enforce an LPIPS loss $\mathcal{L}_{\texttt{lpips}}$ on the estimated clean image $\hat{\mathbf{x}}$ and incorporate an adversarial GAN loss $\mathcal{L}_{\texttt{gan}}$ to refine texture realism. We empirically set $\lambda_{\texttt{FM}} = 1.0$, $\lambda_{\texttt{perc}} = 0.1$, and $\lambda_{\texttt{gan}} = 0.075$. The total reconstruction loss is formulated as $\mathcal{L}_{\texttt{rec}} = \lambda_{\texttt{FM}}\mathcal{L}_{\texttt{FM}} + \lambda_{\texttt{perc}}\mathcal{L}_{\texttt{lpips}} + \lambda_{\texttt{gan}}\mathcal{L}_{\texttt{gan}}$.

Total objective. Finally, the unified tokenizer is optimized by minimizing the weighted sum of the reconstruction, regularization, and alignment objectives defined above:

$$\mathcal{L}_{\texttt{tokenizer}} = \mathcal{L}_{\texttt{rec}} + \mathcal{L}_{\texttt{reg}} + \lambda_{\texttt{dist}}\mathcal{L}_{\texttt{dist}}, \tag{8}$$

where the distillation weight is set to $\lambda_{\texttt{dist}} = 1.0$ by default.

2.2 HYDRA

Built upon the robust representations of HYDRA-TOK, HYDRA represents a native unified framework that integrates understanding and generation within a single parameter space. By leveraging the coherent visual representations from our tokenizer, HYDRA processes visual and textual signals as a unified sequence via a shared autoregressive transformer, employing a specialized dual-head mechanism to reconcile their distinct output modalities.

Unified Input Representation. To achieve the *Unification of Input Representation*, we integrate visual signals into the LLM by treating them as continuous sequences fully compatible with the embedding space. For a given image, we extract its latent representation $\mathbf{H}_{\texttt{bn}}$ from the GSB. We implement a task-aware injection strategy: for flow-based generation, we introduce diffusion noise corresponding to the time-step t ($\mathbf{H}_{\texttt{in}} = \mathbf{H}_{\texttt{bn}} + t\boldsymbol{\epsilon}$); conversely, for understanding, we utilize the clean, deterministic latents to maximize semantic clarity. These processed latents are mapped to the LLM's dimension via a linear projector to yield $\mathbf{U}_{\texttt{vis}}$. Finally, we concatenate them with text embeddings $\mathbf{E}_{\texttt{text}}$ to form a unified input stream:

$$\mathbf{U}_{\texttt{in}} = [\mathbf{E}_{\texttt{text}}, \mathbf{U}_{\texttt{vis}}]. \tag{9}$$

Dual-Head Decoding. The unified sequence $\mathbf{U}_{\texttt{in}}$ is processed by the shared LLM backbone. Following standard UMM practices [3, 1], we apply a causal attention mask on language tokens and a bidirectional attention mask on visual tokens within the LLM decoder layers. To address the *Compatibility of Learning Process* between discrete text prediction and continuous image synthesis, we branch the output of the final transformer layer into two specialized heads. The Language Head (autoregressive) functions as a standard linear classifier, predicting the probability distribution over the text vocabulary based on the unified context: $\mathcal{L}_{\texttt{NTP}} = -\sum \log P(\mathbf{y}_i|\mathbf{y}_{<i}, \mathbf{U}_{\texttt{in}})$. Simultaneously, the Vision Head (diffusion) serves as a continuous regression head modulated by time-step embeddings via AdaLN-Zero [21]. This head operates explicitly on the hidden states corresponding to visual tokens ($\mathbf{H}^{\texttt{vis}}_{\texttt{LLM}}$) to predict the flow velocity:

$$\mathbf{v}_{\texttt{pred}} = \texttt{Head}_{\texttt{flow}}(\texttt{AdaLN}(\mathbf{H}^{\texttt{vis}}_{\texttt{LLM}}, t_{\texttt{emb}})). \tag{10}$$

Unified Training Objective. HYDRA is trained end-to-end via a composite objective:

$$\mathcal{L}_{\texttt{total}} = \mathcal{L}_{\texttt{NTP}} + \mathcal{L}_{\texttt{FM}}. \tag{11}$$

By sharing the LLM backbone while specializing only the input and output layers, HYDRA achieves rational unification.

2.3 Training Recipe

To cultivate the harmonized nature of HYDRA, we implement a three-stage progressive training strategy. We initialize HYDRA-TOK via reconstruction on large-scale corpora combined with semantic self-distillation against a teacher ViT [16]. Subsequent stages incrementally refine the compatibility between understanding and generation (see Appendix D for full data details).

Table 1: **Evaluation on multimodal understanding benchmarks.** # Params. and # Data denote the model size and the volume of image-text pairs used for instruction tuning, respectively. Rows in gray indicate models exceeding 14B parameters.

Models	# Params	# Data	MME-S	AI2D	OCRBench	MMB	RWQA	SEED	MMMU	MMStar	Avg.
Understanding-only Models											
Qwen2.5-VL [22]	3B	1B+	2199.9	81.4	82.8	77.6	65.5	74.0	51.2	56.3	70.9
Qwen2.5-VL [22]	7B	1B+	2312.0	84.3	88.8	82.8	68.4	77.0	58.0	64.1	75.7
LLaVA-1.5 [23]	7B	0.7M	1510.7	55.5	31.8	62.3	54.8	65.8	35.7	33.1	49.1
LLaVA-OV [24]	7B	3.2M	1998.1	81.4	62.2	80.8	66.3	76.7	48.8	61.7	68.7
Composite Unified Multimodal Models											
BLIP3-o [11]	4B	N/A	2161.2	-	-	-	60.4	73.8	46.6	-	-
TokenFlow-XL [25]	14B	N/A	1922.1	-	-	68.9	56.6	72.6	43.2	-	-
SEED-X [9]	17B	N/A	1457.0	70.0	-	70.1	49.1	65.6	35.6	-	-
Smaller Native Unified Multimodal Models											
Show-o [26]	1.3B	0.7M	1345.1	-	-	-	-	51.5	27.4	-	-
JanusFlow [27]	1.3B	1.4M	1560.1	54.2	-	69.4	41.2	70.5	29.3	40.6	-
Janus-Pro [28]	1.5B	10M	1875.1	64.5	-	75.5	52.6	68.3	36.3	43.1	-
Harmon [5]	1.5B	3.2M	1476.2	57.0	11.2	67.1	49.8	67.1	38.9	35.3	47.4
Show-o2 [3]	1.5B	3.2M	1742.3	69.0	24.5	67.4	56.5	65.6	37.1	43.4	53.2
HYDRA	1.5B	3.2M	**1877.9**	**79.4**	**50.8**	**76.8**	**60.7**	**74.5**	**39.1**	**57.0**	**63.2**
Larger Native Unified Multimodal Models											
Ming-UniVision [29]	16B	20M+	2023.0	82.8	72.4	78.5	-	-	40.3	63.7	-
BAGEL [1]	14B	15.2M	2388.0	89.2	73.3	85.0	72.8	78.5	55.3	-	-
MUSE-VL [30]	7B	10M	-	69.8	-	72.1	-	69.1	39.7	49.6	-
Janus-Pro [28]	7B	10M	1927.1	71.3	-	79.2	58.0	72.1	41.0	48.3	-
VILA-U [31]	7B	4M	1845.8	-	-	66.6	-	59.0	32.2	-	-
Show-o2 [3]	7B	3.2M	1920.5	78.6	32.4	79.3	64.7	69.8	48.9	56.6	62.4
HYDRA	7B	3.2M	**2068.6**	**85.1**	**57.7**	**82.4**	**64.7**	**75.5**	**49.4**	**62.3**	**68.9**

Stage I: Unified Representation Alignment. To resolve the representation divergence at the input level, we freeze the LLM and exclusively tune the vision components. Utilizing 100M image-text pairs, this phase aligns the visual latent space with the linguistic domain, ensuring a coherent unified input representation.

Stage II: 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 I). This full-parameter update ensures the *Compatibility of Learning Process*, allowing the two tasks to mutually reinforce each other.

Stage III: High-Quality Instruction Fine-tuning. The final stage focuses on high-fidelity refinement using curated datasets. We employ 3.2M balanced instruction-tuning samples for understanding, alongside 10M aesthetic-filtered images (derived from Stage II) and 6M high-fidelity synthetic images for generation.

3 Experiments

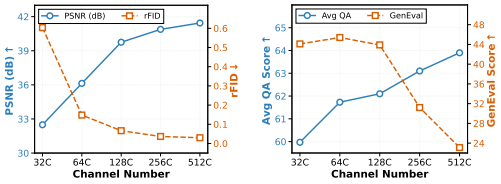
We conduct extensive experiments to evaluate how HYDRA addresses the fundamental challenges in unified multimodal learning identified in the Introduction:

- **RQ1**: What is the sweet spot between high-level semantic abstraction and compact structural synthesis in HYDRA-TOK to achieve a rational unified representation?
- **RQ2**: Can a pure ViT-based tokenizer HYDRA-TOK achieve state-of-the-art reconstruction fidelity?
- **RQ3**: Does the pure ViT architecture HYDRA preserve the coherence of information flow better than VAE-based or decoupled approaches, thereby enhancing multimodal understanding?
- **RQ4**: Does the joint optimization of autoregressive and flow-matching objectives achieve compatibility, enabling state-of-the-art generative performance?

Implementation Details. We evaluate HYDRA at two scales: 1.5B (based on Qwen2.5-1.5B-Instruct [32]) and 7B (Qwen2.5-7B-Instruct). The HYDRA-TOK is initialized with InternViT-2.5 [16] to leverage robust visual priors. The training recipe employs a three-stage progressive pipeline optimized with AdamW [33]. The specific learning rate schedules for each stage are detailed in Tab. 5. For comprehensive experimental details and evaluation metrics regarding all ablation studies, please refer to Appendix D.1.

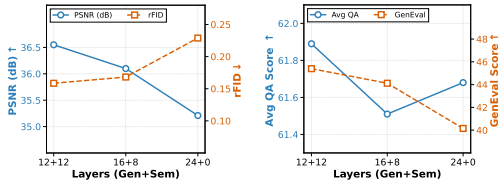
Table 2: **Comprehensive image generation results on GenEval [18], DPG-Bench [19], and WISE [20].** # Data. indicates the number of image-text pairs used for visual generation. † refers to fintuning with GPT-4o distilled synthetic dataset [11]. More Visualization can be founded in Appendix H.2.

Models	# Params	# Data	GenEval			DPG-Bench			WISE		
			Two Obj.	Pos.	Over.	Glob.	Rel.	Over.	Cult.	Space	Over.
Generation-only Models											
SD3-Med [34]	2B	1B+	0.94	0.33	0.74	-	-	84.08	-	-	-
FLUX.1 [Dev] [35]	12B	1B+	0.93	0.68	0.82	82.10	91.10	84.00	0.48	0.62	0.50
Composite Unified Multimodal Models											
Blip3-o† [11]	4B	30M	-	-	0.84	-	-	79.36	-	-	0.50
TokenFlow-XL [25]	14B	60M	0.60	0.16	0.55	87.53	90.30	85.02	-	-	-
SEED-X [9]	17B	158M+	0.58	0.19	0.49	-	-	-	-	-	-
Smaller Native Unified Multimodal Models											
Show-o [26]	1.3B	36M	0.52	0.11	0.53	-	-	67.27	-	-	0.35
Harmon [5]	1.5B	100M	0.86	0.74	0.76	-	-	-	0.38	0.52	0.41
Show-o2 [3]	1.5B	66M	0.86	0.46	0.73	87.53	90.30	85.02	0.33	0.53	0.39
HYDRA	1.5B	100M	**0.95**	**0.76**	**0.86**	**89.65**	**91.59**	**85.51**	**0.49**	**0.64**	**0.49**
Larger Native Unified Multimodal Models											
Ming-UniVision [29]	16B	1B+	0.93	0.92	0.85	-	-	82.12	-	-	-
BAGEL [1]	14B	1.6B	0.95	0.78	0.88	88.94	90.82	85.07	0.44	0.68	0.52
MUSE-VL [30]	7B	24M	0.64	0.25	0.57	-	-	-	-	-	-
Janus-Pro [28]	7B	144M	0.89	0.79	0.80	86.90	89.32	84.19	0.30	0.49	0.35
VILA-U [31]	7B	15M	-	-	-	-	-	-	0.26	0.37	0.31
Show-o2 [3]	7B	66M	0.87	0.52	0.76	89.00	91.81	86.14	0.40	0.58	0.44
HYDRA	7B	100M	**0.97**	**0.81**	**0.86**	**90.22**	**91.83**	**86.40**	**0.52**	**0.68**	**0.53**



(a) Reconstruction (b) Und. vs. Gen.

Figure 4: **Ablation results on different latent channel dimensions.**



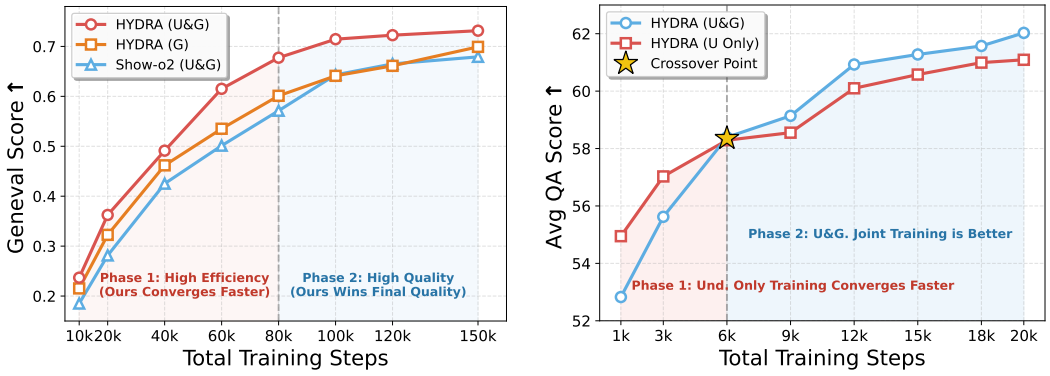
(a) Reconstruction (b) Und. vs. Gen.

Figure 5: **Ablation results on different layer configurations.**

3.1 Main Results

RQ1: Identifying the Sweet Spot for Rational Unification. Achieving a rational unified representation requires identifying the optimal synergy between high-level semantic abstraction and compact structural synthesis within a single feature space. We seek the architectural "sweet spot" where both capabilities are maximized simultaneously, rather than viewing them as competing objectives. As initially determined regarding channel dimensions (referencing Fig. 4), incorporating the GSB to compress the feature space to $C = 64$ represents the representational sweet spot, achieving *Unification of Input Representation* by filtering redundant noise while retaining semantic capacity. Further investigating this balance at the architectural level, Fig. 5 illustrates the impact of different layer allocations. We observe that configurations deviating from a balanced structure fail to achieve optimal synergy. The highly imbalanced $24 + 0$ configuration (pure Gen-ViT) exhibits the poorest performance across reconstruction fidelity (lowest PSNR, highest rFID), generative capability (lowest GenEval Score), and understanding (Avg QA). Similarly, the $16 + 8$ configuration also yields suboptimal results across these metrics compared to the balanced approach. Consequently, the $12 + 12$ configuration clearly emerges as the architectural "sweet spot." It achieves peak performance across all evaluated metrics (highest PSNR, Avg QA, and GenEval Score), effectively harmonizing the depth required for semantic abstraction with the structural stability needed for generation.

RQ2: High-Fidelity Reconstruction. We next verify whether our pure ViT-based HYDRA-TOK can serve as a high-quality visual foundation. Historically, ViTs were deemed unsuitable for dense pixel-level synthesis compared to Convolutional VAEs. On ImageNet-1K under identical 256×256 settings (†), HYDRA-TOK achieves 36.39 PSNR and 0.08 rFID, significantly outperforming the unified baseline UniFlow [36] (33.23 PSNR, 0.26 rFID). At 448×448 resolution, it maintains



(a) Generation. (b) Understanding.

Figure 6: **Analysis of representation-harmonized co-promotion.** (a) For generation, joint training enhances generation efficiency by stabilizing the latent space. (b) For understanding, joint training surpasses single-task baselines after a crossover point, showing that generative constraints refine perceptual precision. Details of each understanding benchmark and experiment setting is shown in Fig. 8 and Appendix D.2

comparable reconstruction quality (31.90 PSNR, 0.12 rFID) against the generative-only FLUX-VAE [35] (32.91 PSNR, 0.08 rFID). This confirms our *functionally progressive design* enables pure ViTs to successfully supersede traditional VAEs.

RQ3: Coherence of Information Flow. We assess the coherence of information flow by evaluating multimodal understanding performance, as shown in Tab. 1. Sequential architectures, such as Show-o2 [3], often compromise information coherence due to the compression bottleneck imposed by VAEs. In contrast, HYDRA maintains a continuous and uncompressed information flow throughout its pure ViT backbone. At the 1.5B scale, HYDRA demonstrates dominant performance with an average score of 63.1, surpassing the strong baseline Show-o2 (53.2) by a significant margin. This advantage is particularly pronounced in fine-grained tasks sensitive to information loss, such as OCRBench [43], where HYDRA achieves 50.8, a score more than double that of Show-o2 (24.5). Scaling up to 7B, HYDRA continues to lead, outperforming Show-o2 on complex reasoning benchmarks like MMStar [44] (62.3 vs. 56.6) and SEED [45] (75.5 vs. 69.8). Most notably, in OCRBench [43], HYDRA preserves high-frequency character details often discarded by VAE-based unified models, scoring 57.7 compared to 32.4 for Show-o2.

Table 3: **Reconstruction results on the ImageNet-1K validation set.** † indicates models trained at a resolution of 448 and resized to 256 for evaluation.

Method	Res.	PSNR ↑	SSIM ↑	rFID ↓
Generative Only Tokenizer				
SD-VAE [37]	256	23.54	0.68	1.22
FLUX-VAE [35]	256	32.74	0.92	0.18
VA-VAE [38]	256	27.96	0.79	0.28
DC-AE [39]	512	26.15	0.71	0.22
FLUX-VAE [35]	448	32.91	0.92	0.08
Unified Tokenizer				
Tokenflow [25]	384	22.73	0.73	0.63
UniTok [4]	256	27.28	0.77	0.41
UniLIP [10]	448	24.62	0.79	0.31
BLIP3-o [11]	256	14.71	0.58	3.18
QLIP [40]	256	23.16	0.63	3.21
UAE [41]	256	29.65	0.88	0.19
RAE [42]	256	18.05	0.50	2.04
UniFlow † [36]	256	33.23	0.96	0.26
HYDRA-TOK †	256	**36.39**	**0.96**	**0.08**
HYDRA-TOK	448	**31.90**	**0.90**	**0.12**

RQ4: Compatibility of Generation Learning Process. We begin by investigating the mechanism behind this compatibility through the training dynamics illustrated in Fig. 6. First, regarding generative harmony (Fig. 6a), we observe that joint training (U&G) consistently outperforms single-task generation (G Only). It not only demonstrates higher convergence efficiency in the early stage (Phase 1) but also achieves superior final fidelity (Phase 2). This validates that the semantic alignment provided by understanding tasks effectively stabilizes the generative latent space, leading to faster and better convergence. Second, regarding understanding harmony (Fig. 6b), a distinct trend appears. While single-task understanding (U Only) learns faster initially, joint training (U&G) overtakes it after a critical *crossover point* (∼6k steps). This phenomenon demonstrates that although generation tasks are harder to optimize initially, their fine-grained structural constraints eventually refine the model's perceptual precision, proving that generation and understanding are mutually reinforcing in our framework.

Supported by this harmonized training dynamic, we further validate the learning compatibility on text-to-image generation benchmarks in Tab. 2. A common failure mode in UMMs is the "tug-of-war" where improving understanding degrades generation. HYDRA breaks this trade-off, demonstrating that joint optimization leads to superior generative performance. At the 1.5B scale, HYDRA establishes a new benchmark for native UMMs, achieving an Overall GenEval score of 0.86 and DPG-Bench score of 85.51, significantly outperforming Show-o2 (0.73 / 85.02). At the 7B scale, HYDRA sets new state-of-the-art records with a GenEval Overall score of 0.86, surpassing both the unified model Ming-UniVision [29] (0.85) and the specialized 12B model FLUX.1 (0.82) [35]. Furthermore, HYDRA achieves an overall score of 0.53 on the WISE, demonstrating robust alignment across diverse cultural and spatial contexts compared to existing native unified models.

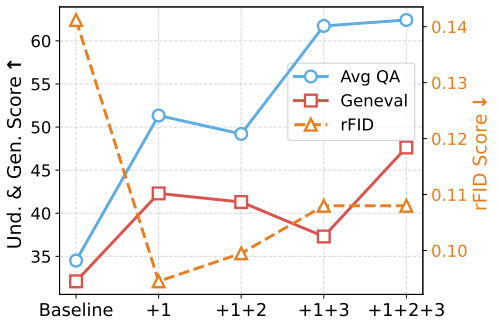


Figure 7: **Ablation of HYDRA-TOK loss.** Baseline utilizes only $\mathcal{L}_{\text{rec}}$. "1": Teacher init; "2": $\mathcal{L}_{\text{reg}}$; "3": $\mathcal{L}_{\text{dist}}$.

3.2 Ablation Analysis

HYDRA-TOK Training Objectives. Fig. 7 validates our tokenizer's training objectives. The baseline, relying solely on reconstruction loss ($\mathcal{L}_{\text{rec}}$), suffers feature collapse and yields suboptimal results. Teacher initialization provides a crucial semantic scaffold, boosting performance, while distillation ($\mathcal{L}_{\text{dist}}$) further enhances comprehension without compromising structure. Ultimately, the complete objective achieves optimal synergy, balancing understanding, generation, and reconstruction to robustly support the unified architecture.

HYDRA Training Stages. Tab. 4 confirms the indispensability of each stage in our progressive training recipe. Omitting Stage I causes a universal performance decline, underscoring its role in initial alignment. Removing Stage II severely degrades generation while impairing understanding, proving it critical for consolidating the unified feature space. Skipping Stage III results in a total loss of instruction-following capabilities for QA tasks. Consequently, the full recipe yields peak performance across all benchmarks.

Table 4: **Ablation study on the effects of the three training stages.**

Stage I	Stage II	Stage III	Avg	GenEval
✗	✓	✓	0.719	0.82
✓	✗	✓	0.683	0.77
✓	✓	✗	-	0.84
✓	✓	✓	**0.744**	**0.86**

4 Conclusion

In this work, we present HYDRA-TOK and HYDRA to reconcile the intrinsic conflict between visual understanding and generation. By employing the Generation-Semantic Bottleneck, HYDRA-TOK functions as a progressive learner that harmonizes structural primitives with semantic abstractions, effectively achieving the *Unification of Input Representation*. This cohesive foundation enables HYDRA to integrate understanding and generation within a single parameter space, satisfying the critical criteria of *Coherence of Information Flow* and *Compatibility of Learning Process*. Our extensive experiments not only establish new state-of-the-art benchmarks but also reveal a fundamental insight: understanding and generation are not competitive objectives but complementary forces that drive mutual enhancement through rational unification. We believe this representation-harmonized paradigm establishes a new standard for native UMMs, offering a promising trajectory toward more versatile and scalable multimodal intelligence.

Acknowledgements

This work was supported by the Zhongguancun Academy (Grant No.s 1800302002).

References

[1] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025.

[2] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 966–12 977.

[3] J. Xie, Z. Yang, and M. Z. Shou, "Show-o2: Improved native unified multimodal models," *arXiv preprint arXiv:2506.15564*, 2025.

[4] C. Ma, Y. Jiang, J. Wu, J. Yang, X. Yu, Z. Yuan, B. Peng, and X. Qi, "Unitok: A unified tokenizer for visual generation and understanding," *arXiv preprint arXiv:2502.20321*, 2025.

[5] S. Wu, W. Zhang, L. Xu, S. Jin, Z. Wu, Q. Tao, W. Liu, W. Li, and C. C. Loy, "Harmonizing visual representations for unified multimodal understanding and generation," *arXiv preprint arXiv:2503.21979*, 2025.

[6] S. Cao, H. Chen, P. Chen, Y. Cheng, Y. Cui, X. Deng, Y. Dong, K. Gong, T. Gu, X. Gu *et al.*, "Hunyuanimage 3.0 technical report," *arXiv preprint arXiv:2509.23951*, 2025.

[7] C. Liao, L. Liu, X. Wang, Z. Luo, X. Zhang, W. Zhao, J. Wu, L. Li, Z. Tian, and W. Huang, "Mogao: An omni foundation model for interleaved multi-modal generation," *arXiv preprint arXiv:2505.05472*, 2025.

[8] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model," *arXiv preprint arXiv:2408.11039*, 2024.

[9] Y. Ge, S. Zhao, J. Zhu, Y. Ge, K. Yi, L. Song, C. Li, X. Ding, and Y. Shan, "Seed-x: Multimodal models with unified multi-granularity comprehension and generation," *arXiv preprint arXiv:2404.14396*, 2024.

[10] H. Tang, C. Xie, X. Bao, T. Weng, P. Li, Y. Zheng, and L. Wang, "Unilip: Adapting clip for unified multimodal understanding, generation and editing," *arXiv preprint arXiv:2507.23278*, 2025.

[11] J. Chen, Z. Xu, X. Pan, Y. Hu, C. Qin, T. Goldstein, L. Huang, T. Zhou, S. Xie, S. Savarese *et al.*, "Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset," *arXiv preprint arXiv:2505.09568*, 2025.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[13] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[14] Z. Liu, W. Ren, H. Liu, Z. Zhou, S. Chen, H. Qiu, X. Huang, Z. An, F. Yang, A. Patel *et al.*, "Tuna: Taming unified visual representations for native unified multimodal models," *arXiv preprint arXiv:2512.02014*, 2025.

[15] Y. Jiao, H. Qiu, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang, "Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3600–3610.

[16] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[17] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

[18] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *Advances in Neural Information Processing Systems*, vol. 36, pp. 52 132–52 152, 2023.

[19] X. Hu, R. Wang, Y. Fang, B. Fu, P. Cheng, and G. Yu, "Ella: Equip diffusion models with llm for enhanced semantic alignment," *arXiv preprint arXiv:2403.05135*, 2024.

[20] Y. Niu, M. Ning, M. Zheng, W. Jin, B. Lin, P. Jin, J. Liao, C. Feng, K. Ning, B. Zhu *et al.*, "Wise: A world knowledge-informed semantic evaluation for text-to-image generation," *arXiv preprint arXiv:2503.07265*, 2025.

[21] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[22] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[23] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[24] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[25] L. Qu, H. Zhang, Y. Liu, X. Wang, Y. Jiang, Y. Gao, H. Ye, D. K. Du, Z. Yuan, and X. Wu, "Tokenflow: Unified image tokenizer for multimodal understanding and generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2545–2555.

[26] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou, "Show-o: One single transformer to unify multimodal understanding and generation," *arXiv preprint arXiv:2408.12528*, 2024.

[27] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, X. 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*, 2025, pp. 7739–7751.

[28] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.

[29] Z. Huang, D. Zheng, C. Zou, R. Liu, X. Wang, K. Ji, W. Chai, J. Sun, L. Wang, Y. Lv *et al.*, "Ming-univision: Joint image understanding and generation with a unified continuous tokenizer," *arXiv preprint arXiv:2510.06590*, 2025.

[30] R. Xie, C. Du, P. Song, and C. Liu, "Muse-vl: Modeling unified vlm through semantic discrete encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24 135–24 146.

[31] Y. Wu, Z. Zhang, J. Chen, H. Tang, D. Li, Y. Fang, L. Zhu, E. Xie, H. Yin, L. Yi *et al.*, "Vila-u: a unified foundation model integrating visual understanding and generation," *arXiv preprint arXiv:2409.04429*, 2024.

[32] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[34] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[35] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser *et al.*, "Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space," *arXiv preprint arXiv:2506.15742*, 2025.

[36] Z. Yue, H. Zhang, X. Zeng, B. Chen, C. Wang, S. Zhuang, L. Dong, K. Du, Y. Wang, L. Wang *et al.*, "Uniflow: A unified pixel flow tokenizer for visual understanding and generation," *arXiv preprint arXiv:2510.10575*, 2025.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[38] J. Yao, B. Yang, and X. Wang, "Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 703–15 712.

[39] A. Shaulov, I. Hazan, L. Wolf, and H. Chefer, "Flowmo: Variance-based flow guidance for coherent motion in video generation," *arXiv preprint arXiv:2506.01144*, 2025.

[40] Y. Zhao, F. Xue, S. Reed, L. Fan, Y. Zhu, J. Kautz, Z. Yu, P. Krähenbühl, and D.-A. Huang, "Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation," *arXiv preprint arXiv:2502.05178*, 2025.

[41] W. Fan, H. Diao, Q. Wang, D. Lin, and Z. Liu, "The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding," *arXiv preprint arXiv:2512.19693*, 2025.

[42] B. Zheng, N. Ma, S. Tong, and S. Xie, "Diffusion transformers with representation autoencoders," *arXiv preprint arXiv:2510.11690*, 2025.

[43] Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai, "Ocrbench: on the hidden mystery of ocr in large multimodal models," *Science China Information Sciences*, vol. 67, no. 12, p. 220102, 2024.

[44] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 056–27 087, 2024.

[45] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[46] Y. Zhao, W. Lin, Z. Zhang, M. Yang, J. Chen, W. Zhang, Z. Zhong, and L. Bo, "Unicom: Unified multimodal modeling via compressed continuous semantic representations," *arXiv preprint arXiv:2603.10702*, 2026.

[47] J. Hu, S. Zhao, Q.-G. Chen, X. Qiu, J. Liu, Z. Xu, W. Luo, K. Zhang, and Y. Lu, "Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images," *arXiv preprint arXiv:2511.07222*, 2025.

[48] J. Li, S. Zhou, L. Guo, X. Qiu, L. Xu, D. Qu, T. Long, C. Fan, M. Li, H. Fan *et al.*, "Uniface: A unified fine-grained face understanding and generation model," *arXiv preprint arXiv:2503.08120*, 2025.

[49] S. Tu, Q. Dai, Z.-Q. Cheng, H. Hu, X. Han, Z. Wu, and Y.-G. Jiang, "Motioneditor: Editing video motion via content-aware diffusion," in *CVPR*, 2024.

[50] S. Tu, Q. Tian, Z. Yang, Y. Wu, X. Han, W. Kong, J. Xiong, J.-W. Zhang, Z. Zhong, L. Bo *et al.*, "Baton: Explicit semantic blueprints for joint video-audio generation," *arXiv preprint arXiv:2605.25195*, 2026.

[51] J. Hu, T. Chen, Y. Pan, X. Qiu, J. Liu, X. Cao, Z. Zhu, G. Chen, H. Ye, B. Wang *et al.*, "Learning to retrieve: Internalizing memory retrieval for video world models," *arXiv preprint arXiv:2610.11444*, 2026.

[52] X. Pan, S. N. Shukla, A. Singh, Z. Zhao, S. K. Mishra, J. Wang, Z. Xu, J. Chen, K. Li, F. Juefei-Xu *et al.*, "Transfer between modalities with metaqueries," *arXiv preprint arXiv:2504.06256*, 2025.

[53] S. Wu, Z. Wu, Z. Gong, Q. Tao, S. Jin, Q. Li, W. Li, and C. C. Loy, "Openuni: A simple baseline for unified multimodal understanding and generation," *arXiv preprint arXiv:2505.23661*, 2025.

[54] M. Shi, H. Wang, W. Zheng, Z. Yuan, X. Wu, X. Wang, P. Wan, J. Zhou, and J. Lu, "Latent diffusion model without variational autoencoder," *arXiv preprint arXiv:2510.15301*, 2025.

[55] J. Yao, Y. Song, Y. Zhou, and X. Wang, "Towards scalable pre-training of visual tokenizers for generation," *arXiv preprint arXiv:2512.13687*, 2025.

[56] S. Tu, Y. Pan, Y. Huang, X. Han, Z. Xing, Q. Dai, C. Luo, Z. Wu, and Y.-G. Jiang, "Stableavatar: Infinite-length audio-driven avatar video generation," *arXiv preprint arXiv:2508.08248*, 2025.

[57] Z. Yang, S. Tu, L. Zhang, Q. Dai, Y.-G. Jiang, and Z. Wu, "Arcflow: Unleashing 2-step text-to-image generation via high-precision non-linear flow distillation," *arXiv preprint arXiv:2602.09014*, 2026.

[58] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[59] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," *arXiv preprint arXiv:2305.10355*, 2023.

[60] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[61] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. 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*, 2024, pp. 9556–9567.

[62] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *European conference on computer vision*. Springer, 2016, pp. 235–251.

[63] T. Shen, J. Yu, D. Zhou, D. Li, and E. Barsoum, "E-mmdit: Revisiting multimodal diffusion transformer design for fast image synthesis under limited resources," *arXiv preprint arXiv:2510.27135*, 2025.

[64] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[65] K. Sun, J. Pan, Y. Ge, H. Li, H. Duan, X. Wu, R. Zhang, A. Zhou, Z. Qin, Y. Wang *et al.*, "Journeydb: A benchmark for generative image understanding," *Advances in neural information processing systems*, vol. 36, pp. 49 659–49 678, 2023.

[66] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211 – 252, 2014.

[67] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 092–27 112, 2023.

[68] M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim, "Coyo-700m: Image-text pair dataset," https://github.com/kakaobrain/coyo-dataset, 2022.

[69] X. Li, F. Zhang, H. Diao, Y. Wang, X. Wang, and L. Duan, "Densefusion-1m: Merging vision experts for comprehensive multimodal perception," *Advances in Neural Information Processing Systems*, vol. 37, pp. 18 535–18 556, 2024.

[70] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[71] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[72] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *ArXiv*, vol. abs/2102.12092, 2021.

[73] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[74] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[75] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

A Related Works

A.1 Unified Multimodal Models

The pursuit of a singular architecture for both visual understanding and generation has led to the evolution of Unified Multimodal Models (UMMs) [1, 46–48], broadly categorized into composite and native architectures. Complementary generation-focused studies explore various application [49–51].

Composite UMMs such as MetaQuery [52], BLIP-3o [11], OpenUni [53] and Unilip [10] achieve capability extension by integrating specialized, pre-trained experts (e.g., separate LLMs and Diffusion models) via lightweight adapters. While effective for rapid deployment, their reliance on frozen backbones restricts deep inter-modal interaction, resulting in a shallow unification where understanding and generation remain functionally decoupled.

Native UMMs strive for deeper integration by training holistically within a shared parameter space. However, as analyzed in Fig. 2, existing methods are constrained by a fundamental representation trilemma: (i) Decoupled Architectures, such as the Janus series [2, 27], Bagel [1] and UniToken [15], mitigate task interference by utilizing separate or dual encoders for vision and language. While UniToken concatenates discrete and continuous tokens to bridge modalities, this dual-encoder design inherently sacrifices *Unification of Input Representation*, leading to parameter redundancy, computational inefficiency due to sequence inflation, and severed task synergy. (ii) Sequential Architectures, including Show-o2 [3] and TUNA [14], attempt to unify inputs by stacking semantic encoders atop a VAE. However, the aggressive spatial downsampling required by the VAE acts as a lossy information bottleneck, disrupting the *Coherence of Information Flow* and irreversibly diluting the fine-grained structural signals needed for generation. (iii) Shared Architectures, such as Unitok [4], Transfusion [8], and Harmon [5], attempt to unify tasks but struggle with the *Compatibility of Learning Process*. On one hand, discrete approaches like Unitok [4] suffer from quantization errors inherent in codebooks, which fundamentally conflict with the high-fidelity requirements of synthesis. On the other hand, continuous approaches like Transfusion [8] and Harmon [5] employ VAE latent features for both tasks; while these representations are adequate for generation, they lack the high-dimensional semantic abstractions required for perception, resulting in suboptimal performance on complex understanding tasks. In contrast, our HYDRA achieves rational unification by resolving these conflicts through a harmonized, continuous representation.

A.2 Unified Tokenizers and Representations

Recent research has shifted from discrete to continuous unified representations to bridge the gap between semantic abstraction and pixel-level fidelity. To circumvent traditional convolutional or pre-trained VAEs, paradigms such as RAE [42] and SVG [54] employ frozen, pre-trained semantic encoders. While this accelerates convergence, the inherent loss of low-level features compromises reconstruction fidelity. Although UniFlow [36] attempts to recover structural details via self-distillation, it relies on dual-stream features, failing to achieve a truly unified, single-stream information flow. Similarly, QLIP [40] seeks to unify representations by aligning visual tokens with textual semantics; however, this heavy reliance on linguistic alignment imposes a semantic bottleneck, filtering out the high-frequency textural details necessary for photorealistic synthesis. Furthermore, while VTP [55] improves generative scaling by jointly optimizing contrastive and reconstruction losses, its low-dimensional latent space is tailored for generation and lacks the representational capacity required for complex multimodal perception. Complementary to representation learning, Some works [56, 57] improve sampling efficiency through non-linear flow distillation for two-step text-to-image generation. Distinctly, our HYDRA-TOK employs a functionally progressive pure ViT backbone to actively bridge the representation gap. This design ensures a coherent single-stream flow, eliminating quantization errors while simultaneously satisfying the conflicting demands of high-fidelity synthesis and deep semantic understanding.

B Limitations and Future Work

Despite establishing promising benchmarks, our current implementation faces limitations in scale that define future research directions. A primary constraint is the 300M parameter HYDRA-TOK, which creates a potential bottleneck for encoding intricate details in fine-grained tasks. Furthermore, relying

on a 7B parameter model trained on only 100M image-text pairs restricts generalizability and world knowledge absorption compared to state-of-the-art foundation models. Consequently, our immediate future work will center on significantly scaling both the model architecture and the training dataset size to address these gaps and unlock robust capabilities across diverse multimodal tasks.

C Broader Impacts

As HYDRA introduces state-of-the-art 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.

D Training details

HYDRA-TOK is training in two progressive stages to effectively balance representation learning and generative quality. Stage 1: Foundation Training. In the initial stage, we focus on establishing the foundational capabilities of the tokenizer through joint reconstruction and distillation objectives. We utilize a large-scale composite dataset consisting of ImageNet-1.2M, CC-12M, and SAM-10M. During this phase, the entire tokenizer (both encoder and decoder) is trained end-to-end. We optimize the model for a total of 300k iterations with a global batch size of 256. The learning rate is set to $2e-5$. Stage 2: Decoder Refinement. In the second stage, we shift our focus to refining the fidelity of the generated images. We freeze the parameters of the pre-trained encoder to preserve the learned semantic representations and exclusively fine-tune the decoder. In this phase, the distillation objective is discarded. Instead, we incorporate adversarial training (GAN loss) alongside the reconstruction loss to enhance perceptual quality. The learning rate is reduced to $1e$-5, and the training proceeds for an additional 100k iterations.

HYDRA training proceeds through a progressive three-stage curriculum, with specific hyperparameters detailed in Tab. 5. Across all three stages, we maintain a constant input image resolution of 448×448 and a global batch size of 1024. We employ varying learning rates (LRs) tailored to different model components and training phases. In Stage I, the Vision Head and Sem-ViT are trained with LRs of 10^{-4} and 5×10^{-5}, respectively. Progressing to Stages II and III, the Vision Head LR is reduced to 5×10^{-5}, while the LLM and Sem-ViT are trained jointly at a lower rate of 2×10^{-5}. The training strategy also involves shifting data mixing ratios (Text Understanding : Image Understanding : Image Generation) and durations: Stage I utilizes a data ratio of 0:1:3 for 150K steps; Stage II adjusts the ratio to 0:2:2 for 100K steps; and Stage III concludes with a ratio of 1:2:2 for 20K steps.

Table 5: **Training details and computational cost of our HYDRA.** The total training of the UMM stages takes ∼72h on 128 H20 GPUs. The HYDRA-TOK pre-training takes an additional 24h on 64 H20 GPUs (totaling ∼96h). †Data Ratio denotes Text Understanding : Image Understanding : Image Generation.

Setting	Stage I	Stage II	Stage III
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}
Image Size	448	448	448
Batch Size	1024	1024	1024
Data Ratio†	0:1:3	0:2:2	1:2:2
Hardware	128 H20 GPUs	128 H20 GPUs	128 H20 GPUs
Training Step	150K	100K	20K
Time Cost	∼30h	∼30h	∼12h

D.1 Ablation study training details

In our ablation studies, the evaluation covers three core capabilities with specific setups: (i) Multi-modal Understanding: Following the LLaVA-1.5 training protocol [58], we report the average score (Avg QA) across POPE [59], MMBench [60], MMMU [61], AI2D [62], and RealWorldQA. (ii) Image Generation: We adapt the Nitro-E framework [63] by replacing its VAE with our HYDRA-TOK. We train this setup for 15k iterations using the SAM-1B [64] and JourneyDB [65] datasets and evaluate performance on GenEval. (iii) Image Reconstruction: We train on the ImageNet-1k (1.2M) dataset [66] for 150k iterations and assess quality using rFID.

D.2 Representation-harmonized co-promotion experiment training details

To validate the mutual benefits of our harmonized representation, we conducted comparative experiments on both generation and understanding tasks. For generation, we trained both Show-o2 and our HYDRA on the same 100M dataset for 150K steps, comparing the GenEval performance of joint training against a generation-only baseline across different training stages. For understanding, we performed Stage III training for 20K steps, comparing joint training with an understanding-only approach. We evaluated the models on general understanding (SEED-Bench), reasoning (MM-Star, MMMU, MMBench, AI2D), and OCR-related tasks (OCRBench). Detailed settings for each benchmark are illustrated in Fig. 8.

D.3 Training data

D.3.1 Image Understanding

Stage I. For the initial pre-training stage, we utilized a massive dataset comprising approximately 22M images paired with captions. These data were sourced from a diverse collection of large-scale datasets, including ImageNet [66], DataComp-1B [67], COYO-700M [68], SA-1B [64], and DenseFusion [69].

Stage II. In Stage II, we continued to utilize the comprehensive collection of data from the sources employed in Stage I (ImageNet [66], DataComp-1B [67], COYO-700M [68], SA-1B [64], and DenseFusion [69]) for further training.

Stage III. For the final high-quality fine-tuning stage, we directly utilized the refined 3.2M instruction-tuning data from LLaVA-One-Vision [70] to maximize image understanding performance.

D.3.2 Image Generation

Stage I. For the initial generation training stage, we curated a large-scale dataset totaling approximately 78M samples. This comprises SA-1B [64], CC-12M [71], 1M images from DenseFusion [69], and a caption dataset consisting of 55M real images.

Stage II. Stage II expands upon the data used in Stage I. In addition to SA-1B [64], CC-12M [71], the 55M real images, and DenseFusion-1M [69], we incorporated 10M synthetic images generated through Flux distillation, following the methodology described in [63].

Stage III. For high-quality generation fine-tuning, we utilized the JourneyDB dataset [65] and DenseFusion-1M [69]. Additionally, we refined the 10M Flux-distilled data from Stage II through filtering, selecting a high-quality subset of 5M images for this final stage.

E Additional main results

F Evaluation Details

F.1 Multi-modal Understanding Benchmarks

To comprehensively evaluate the perception and reasoning capabilities of HYDRA, we employ eight diverse benchmarks covering general understanding, expert knowledge, and fine-grained visual

Table 6: **Detailed image generation results on the GenEval benchmark.** † refers to fintuning with GPT-4o distilled dataset [11].

Models	Size	Single Object	Two Objects	Count	Colors	Position	Color Attribute	Overall
Generation-only Models								
SD3-Med [34]	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1 [Dev] [35]	12B	0.98	0.93	0.75	0.93	0.68	0.65	0.82
DALL-E 3 [72]	-	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Composite Unified Multimodal Models								
TokenFlow-XL [25]	14B	0.95	0.60	0.41	0.81	0.16	0.24	0.55
SEED-X [9]	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
MetaQuery-XL [52]	7B	-	-	-	-	-	-	0.80
Blip3-o † [11]	8B	-	-	-	-	-	-	0.84
Smaller Native Unified Multimodal Models								
Show-o [26]	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Harmon [5]	1.5B	0.99	0.86	0.66	0.85	0.74	0.48	0.76
Show-o2 [3]	1.5B	0.99	0.86	0.55	0.86	0.46	0.63	0.73
HYDRA	1.5B	**1.00**	**0.95**	**0.75**	**0.91**	**0.76**	**0.77**	**0.86**
Larger Native Unified Multimodal Models								
Ming-UniVision [29]	16B	1.00	0.93	0.59	0.93	0.92	0.70	0.85
BAGEL [1]	14B	0.98	0.95	0.84	0.95	0.78	0.77	0.88
MUSE-VL [30]	7B	0.98	0.64	0.54	0.72	0.25	0.31	0.57
Janus-Pro [28]	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Show-o2 [3]	7B	1.00	0.87	0.58	0.92	0.52	0.62	0.76
HYDRA	7B	**1.00**	**0.97**	**0.68**	**0.91**	**0.81**	**0.80**	**0.86**

Table 7: **Detailed image generation results on the DPG-Bench benchmark [19].**

Models	Size	Global	Entity	Attribute	Relation	Other	Overall
Generation-only Models							
FLUX.1 [Dev] [35]	12B	82.10	89.50	88.70	91.10	89.40	84.00
Smaller Native Unified Multimodal Models							
Show-o2 [3]	1.5B	87.53	90.38	91.34	90.30	91.21	85.02
HYDRA	1.5B	**89.65**	**91.62**	**90.24**	**91.59**	**90.88**	**85.51**
Larger Native Unified Multimodal Models							
Janus-Pro [28]	7B	86.90	88.90	89.40	89.32	89.48	84.19
BAGEL [1]	14B	88.94	90.37	91.29	90.82	88.67	85.07
Show-o2 [3]	7B	89.00	91.78	89.96	91.81	91.64	86.14
HYDRA	7B	**90.22**	**91.88**	**89.92**	**91.83**	**90.93**	**86.40**

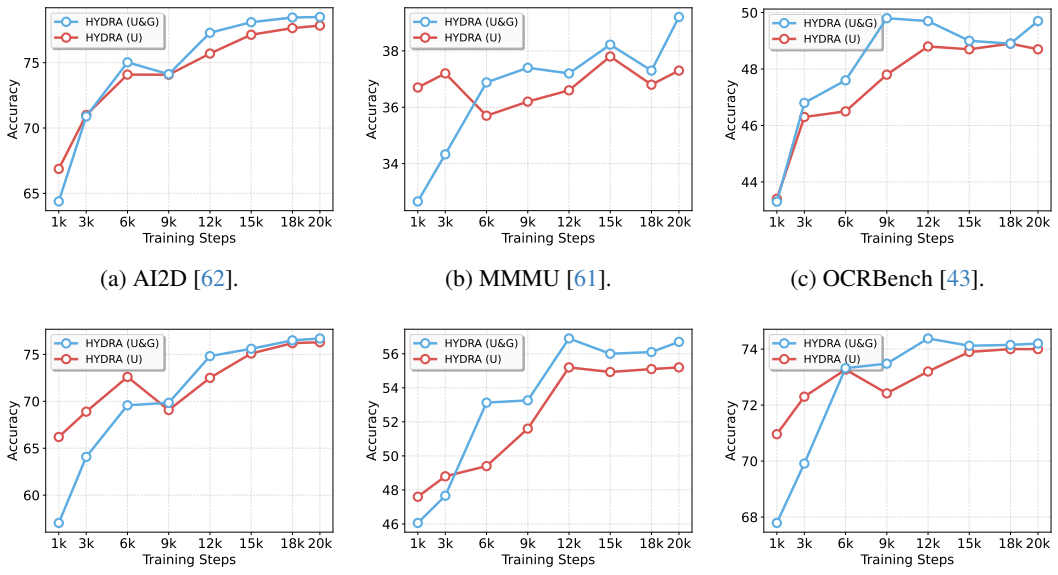
perception. We utilize MME [74] to measure comprehensive perception and cognition, reporting MME-S as the aggregate sum of both scores. MMBench (MMB) [60] is employed for its robust circular evaluation strategy, assessing fine-grained perception and logical reasoning. To ensure rational visual dependency, we evaluate on MMStar [44], a curated benchmark that filters out samples solvable by text alone. Additionally, SEED-Bench (SEED) [45] serves as a large-scale testbed for generative and discriminative comprehension across spatial and temporal dimensions. For high-level discipline-specific reasoning, we employ MMMU [61], a massive multi-discipline benchmark demanding expert-level reasoning across fields such as art, science, and engineering. This is complemented by AI2D [62], a specialized dataset focused on understanding and answering questions about scientific diagrams and textbook illustrations. To assess text-centric visual understanding, we use OCRBench [43], a dedicated benchmark for optical character recognition tasks. Furthermore, RealWorldQA (RWQA) is utilized to evaluate spatial reasoning and physical understanding in diverse real-world environments.

Table 8: **Detailed image generation results on the WISE benchmark [20].**

Models	Size	Culture	Time	Space	Biology	Physics	Chemistry	Overall
Generation-only Models								
FLUX.1 [Dev] [35]	12B	0.48	0.58	0.62	0.42	0.51	0.35	0.50
SD3.5-Large [34]	8B	0.44	0.50	0.58	0.44	0.52	0.31	0.46
Composite Unified Multimodal Models								
Blip3-o-4B [11]	4B	-	-	-	-	-	-	0.50
Smaller Native Unified Multimodal Models								
Harmon [5]	1.5B	0.38	0.48	0.52	0.37	0.44	0.29	0.41
Show-o2 [3]	1.5B	0.33	0.41	0.53	0.35	0.49	0.31	0.39
HYDRA	1.5B	**0.49**	**0.47**	**0.64**	**0.42**	**0.59**	**0.34**	**0.49**
Larger Native Unified Multimodal Models								
BAGEL [1]	14B	0.44	0.55	0.68	0.44	0.60	0.39	0.52
VILA-U-7B [31]	7B	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Janus-Pro-7B [28]	7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3-Gen-8B [73]	8B	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Show-o2 [3]	7B	0.40	0.45	0.58	0.39	0.53	0.34	0.44
HYDRA	7B	**0.52**	**0.53**	**0.68**	**0.47**	**0.58**	**0.38**	**0.53**

Table 9: Summary of licenses for existing assets used in this work.

Asset Name	Asset Type	License
Qwen2.5 [32]	Pre-trained Model	Apache License 2.0
InternViT [16]	Pre-trained Model	Apache License 2.0
ImageNet-1K [66]	Dataset	Custom (Non-commercial research)
SA-1B [64]	Dataset	Custom (Research use only)



(a) AI2D [62]. (b) MMMU [61]. (c) OCRBench [43].

(d) MMBench [60]. (e) MMStar [44]. (f) SEED-Bench [45].

Figure 8: Performance comparison on various benchmarks across different training steps. The benchmarks include general perception (SEED-Bench), reasoning (MMStar, MMMU, MMBench, AI2D), and OCR-related tasks (OCRBench).

G Licenses for Existing Assets

In our research, we utilize several open-source pre-trained models and datasets. We strictly adhere to their respective licenses and terms of use. The explicit licensing terms for the primary assets used in this paper are summarized in Table 9.

H Qualitative comparison

H.1 Image Reconstruction

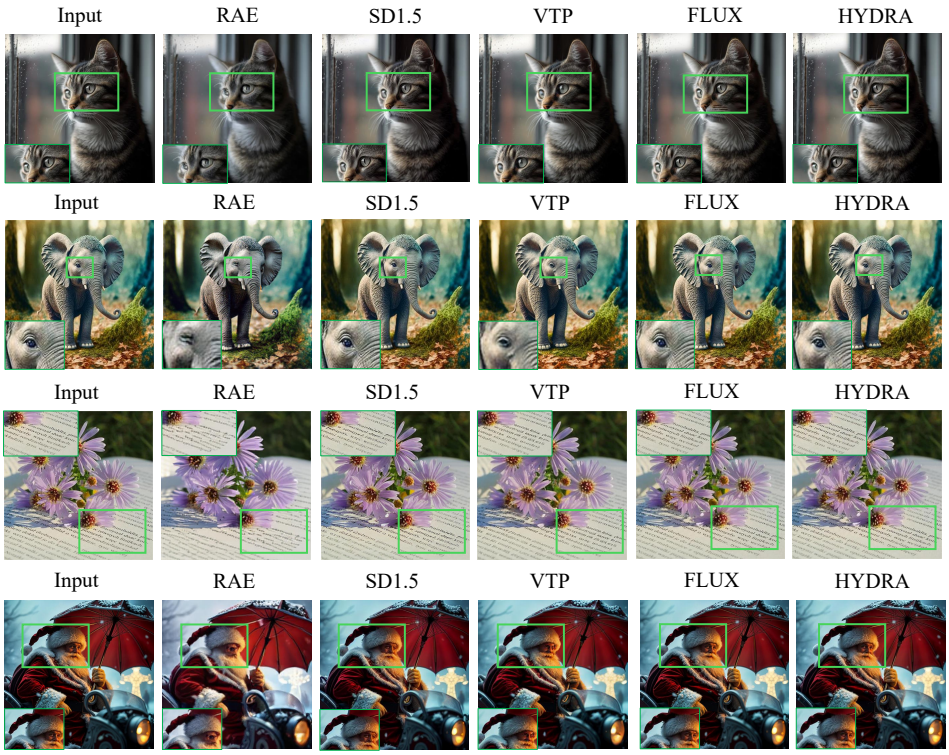


Figure 9: **Qualitative comparison of image reconstruction.** We compare visual results from our HYDRA between RAE [42], SD1.5 [75], VTP [55], and FLUX [35].

H.2 Image Generation



Figure 10: **Qualitative comparison of text-to-image generation.** We compare visual generation between Harmon [5], Show-o2 [3], Bagel [1], Janus-pro [2], and Blip-3o [11].

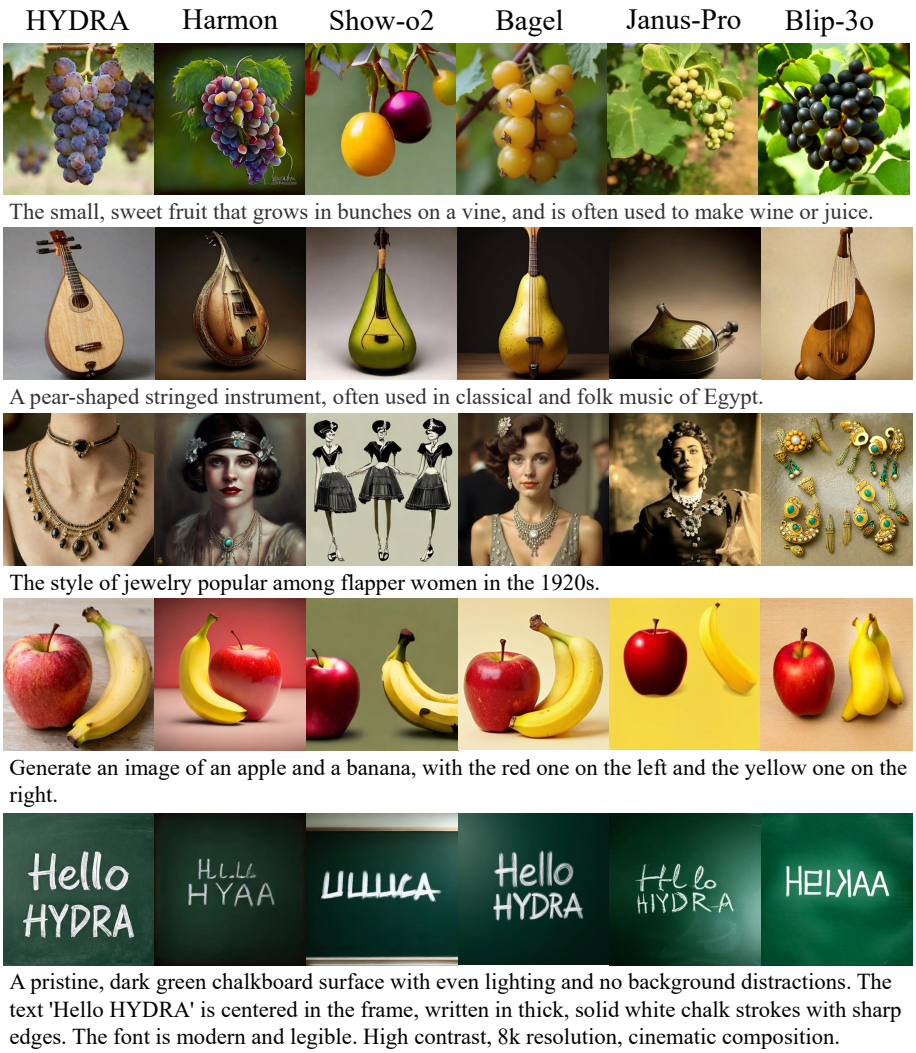


Figure 11: **Qualitative comparison of text-to-image generation.** We compare visual generation results from our HYDRA between Harmon [5], Show-o2 [3], Bagel [1], Janus-pro [2], and Blip-3o [11].

H.3 Multimodal Image Understanding

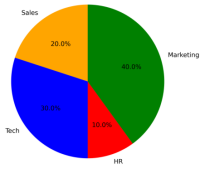
Prompt: Generate the latex code of this formula.

$$A_n = a_0 \left[1 + \frac{3}{4} \sum_{k=1}^{n} \left(\frac{4}{9} \right)^k \right]$$

Response:

```
\[
A_n = a_0 \left[ 1 + \frac{3}{4} \sum_{k=1}^{n}
\left( \frac{4}{9} \right)^{k} \right]
\]
```

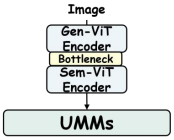
Prompt: Generate the matplotlib.pyplot code for this plot.



Response:

```
import matplotlib.pyplot as plt
labels = ['Sales', 'Tech', 'HR', 'Marketing']
sizes = [20,30,10,40]
colors = ['orange', 'blue', 'red', 'green']
fig, ax = plt.subplots()
ax.pie(sizes, labels=labels, colors=colors,
autopct='%1.1f%%', startangle=90)
ax.axis('equal')
plt.show()
```

Prompt: Describe this image and the algorithm within it.



Response:

The image appears to be a simplified diagram or flowchart representing a conceptual model, possibly related to machine learning or artificial intelligence. It consists of three main components stacked vertically, each labeled with text and accompanied by a smiley face emoji to the left.

At the top is a rectangle labeled "**Gen-ViT Encoder**," which suggests it could be a genetic variant of a **Vision Transformer (ViT)** encoder, a type of neural network used for image recognition tasks. The term "**Gen**" might imply a generative aspect or a genetic algorithm component.

Below the "**Gen-ViT Encoder**" is a smaller rectangle labeled "**Bottleneck.**" This term is often used in machine learning to describe a point in a neural network where the dimensionality of the data is reduced, typically to force the network to learn more abstract features.

The bottom rectangle is labeled "**Sem-ViT Encoder**," which could indicate a semantic variant of a Vision Transformer encoder. This might be referring to a semantic encoder that processes and interprets the meaning or context of visual data.

Beneath these components is a larger rectangle labeled "**UMMs**," which could stand for "**Unified Multimodal Models**" or a similar term, suggesting that these encoders are part of a larger framework or system that integrates different models or processes.

Figure 12: **Qualitative results on multimodal image Understanding.** First, HYDRA exhibits superior proficiency in transforming complex visual data into structured text. From accurately generating LaTeX and Python code for formulas and charts to extracting fine-grained details from concept maps, the results highlight its powerful comprehension of diverse and abstract visual inputs.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 main claims regarding the performance and architectural unification of HYDRA and HYDRA-TOK are comprehensively supported by the extensive experiments and ablations in Section 3.

   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: The authors explicitly discuss the limitations regarding current model scale constraints and training data size limitations in Appendix B.

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 is primarily empirical and architectural, focusing on multimodal model design rather than presenting formal theoretical 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: Comprehensive training details, including learning rates, data mixtures, batch sizes, and stage-by-stage model configurations, are explicitly provided in Section 2.3 and Appendix D.

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: [Yes]

   Justification: The complete source code, pre-trained models, and data preparation scripts will be fully open-sourced upon the acceptance of this paper.

   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: Detailed training settings, optimizers, learning rate schedules, and dataset splits are thoroughly documented in Appendix D and Table 6.

   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: The paper reports absolute metric scores on standard massive benchmarks as is customary for large multimodal foundation models, but does not include error bars due to the prohibitive computational cost of executing multiple full pre-training runs.

   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: Detailed computational resource information, including the specific hardware configurations (H20 GPUs) and execution times for each training stage, is explicitly provided in the training details/appendix section.

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 proposes a generalized unified multimodal architecture and adheres to standard computational research ethics without violating the NeurIPS Code of Ethics.

   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: We discuss potential risks of text-to-image generation (e.g., deepfakes, disinformation) and possible mitigations (e.g., watermarking, gated release) in Appendix B (Broader Impacts).

    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: [N/A]

    Justification: The paper focuses on evaluating the architectural design and benchmark performance; as it has not yet publicly released the generated models, specific release safeguards are not discussed.

    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: We properly cite all existing datasets and foundation models used in the paper (e.g., ImageNet, SA-1B, Qwen2.5). The specific licensing terms and copyright information for each individual asset are explicitly listed and summarized in Appendix G (Licenses for Existing Assets).

    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 introduces new model architectures and progressive training recipes but does not introduce new standalone datasets or external assets for public release.

    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 research exclusively evaluates models on automated, standard benchmarks and does not involve any crowdsourcing or human subject experiments.

    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 human subjects or direct participant studies, rendering IRB approval not applicable.

    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: [N/A]

    Justification: Large Language Models (e.g., Qwen) are utilized purely as the foundational backbone for the proposed multimodal architecture, representing a standard and expected paradigm in this research field rather than a non-standard methodology requiring explicit declaration.

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.