LET VISION-LANGUAGE MODELS GENERATE IMAGES THROUGH THEIR REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

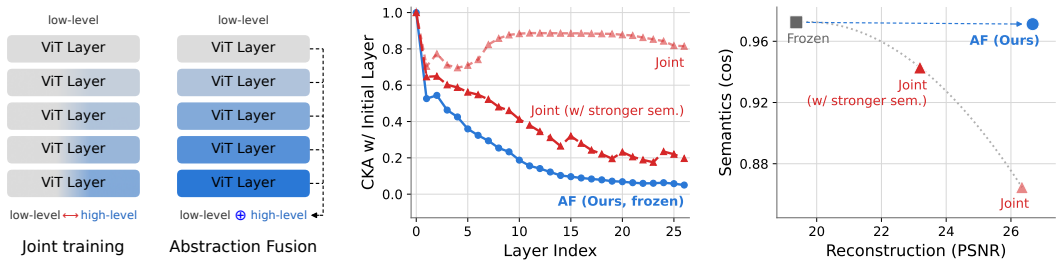
Recent unified multimodal models (UMMs) aim to combine visual understanding and generation within a single model, yet effectively integrating the two capabilities remains challenging. Leveraging pretrained vision-language models (VLMs) offers a way to build on already learned multimodal knowledge, but making this knowledge effectively accessible to visual generation remains nontrivial. We find that a central challenge lies in the heterogeneous visual representations used for understanding and generation. To address this challenge, we introduce *Integral*, which converts a pretrained VLM into a UMM by constructing a shared visual representation space across the two capabilities. *Integral* implements this design through two complementary components. Abstraction Fusion constructs shared visual tokens by aggregating hierarchical representations from the frozen vision encoder, while Block-wise MoT introduces a lightweight generative stream that interacts with the VLM only at block boundaries. Our approach preserves nearly all of the pretrained VLM's understanding capability while extending it to strong visual generation. Controlled comparisons further show that sharing the VLM representation space improves generative learning beyond what can be explained by reconstruction quality alone. Our results indicate that pretrained VLM representations already contain knowledge useful for visual generation, and that preserving their representation hierarchy enables this knowledge to be used without sacrificing visual understanding.

1 INTRODUCTION

Unifying multimodal understanding and generation within a single model has received growing attention in multimodal learning (OpenAI, 2024; Team, 2024; Chen et al., 2025b; Wang et al., 2024). Such unification offers the potential for knowledge learned from one capability to benefit the other. (OpenAI, 2024). However, jointly optimizing these two capabilities remains challenging, with prior work observing performance trade-offs between understanding and generation (Ye et al., 2026; Liu et al., 2026a). To mitigate this interference, recent UMMs often introduce substantial generation-specific capacity to separate generation from the understanding pathway (Liang et al., 2025; Pan et al., 2025; Deng et al., 2025; Shi et al., 2025). Yet such specialization can limit the extent to which visual generation leverages capabilities already learned through multimodal understanding (Liu et al., 2026a; Kang et al., 2026; Wu et al., 2026b).

Prior work has explored how to better exploit capabilities learned through multimodal understanding for visual generation through various training strategies and architectural designs. Some approaches seek to align the two capabilities through training strategies or additional components (Liu et al., 2026a; Kang et al., 2026; Liu et al., 2026b). These approaches require carefully designed optimization strategies to align the two learning objectives. Other approaches explore unified visual representations for multimodal understanding and generation (Wu et al., 2025b; Yue et al., 2026; Yang et al., 2026), achieving deeper integration between understanding and generation (Kang et al., 2026). Such unified representations require balancing semantic understanding and visual reconstruction, creating a trade-off between preserving understanding capabilities and achieving high-quality generation (Wu et al., 2025c; Lin et al., 2025; Huang et al., 2025; Liu et al., 2026c). This raises the question of whether representations learned by a pretrained vision-language models (VLMs) for multimodal understanding can also be used for visual generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Joint vs. AF (Ours) (b) Layer-wise similarity (c) Semantics vs. Reconstruction

Figure 1: **Propagation bottleneck and Abstraction Fusion (AF).** We compare tokenizers built on the same Qwen3.5-9B ViT and trained for 20k steps. **Joint** fine-tunes the encoder for semantics and reconstruction with $\lambda_{\text{sem}} = 1$ (**stronger sem.**: $\lambda_{\text{sem}} = 10$), while **AF** freezes the encoder and fuses features across its hierarchy. (a) Illustration of the propagation-bottleneck hypothesis. **Joint** requires reconstruction-relevant information to persist through deeper layers, whereas **AF** directly fuses features across the frozen encoder hierarchy. (b) CKA similarity between each layer and the initial-layer representation. **Joint** keeps deeper layers more similar to early representations, while stronger semantic supervision reduces this effect. (c) **Joint** exhibits a trade-off between semantic preservation (cosine similarity) and reconstruction fidelity (PSNR), whereas **AF** maintains semantic similarity close to the frozen baseline while substantially improving reconstruction.

In this work, we propose *Integral*, a unified multimodal modeling approach that extends a pretrained vision-language model with visual generation while preserving its understanding capability. Large-scale pretrained vision-language models already provide strong multimodal understanding capabilities (Bai et al., 2025a; Clark et al., 2026; Wang et al., 2025b; Qwen Team, 2026). Our approach builds on these capabilities by constructing a visual representation space shared by understanding and generation. *Integral* realizes this design through two complementary components. We first propose Abstraction Fusion (AF), a visual tokenization method that constructs shared visual tokens by aggregating representations across the frozen vision encoder hierarchy, enabling reconstruction without modifying the pretrained hierarchy. Building on this, we introduce Block-wise MoT, which adds a lightweight generative stream that receives conditioning from the frozen VLM only at block boundaries.

Contributions. We introduce *Integral*, which extends a pretrained VLM with strong visual generation using only 0.89B generation-specific parameters while preserving nearly all of its understanding capability. *Integral* achieves strong performance across compositional generation, knowledge-intensive generation, and image editing that requires multimodal understanding. Controlled comparisons show that sharing the VLM representation space improves generative learning beyond what can be explained by reconstruction quality alone.

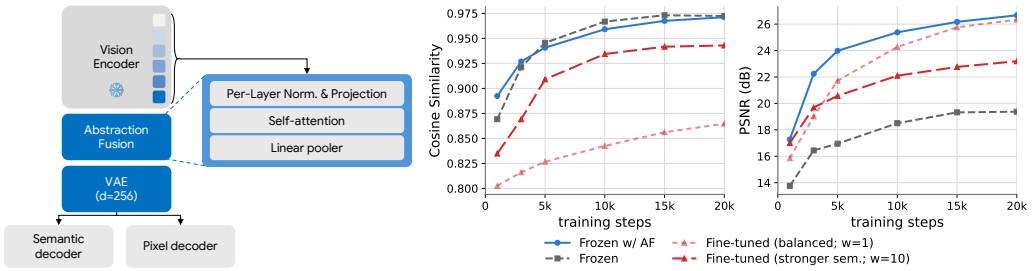
2 Method

We introduce *Integral*, a unified multimodal modeling approach that leverages a pretrained VLM to support both visual understanding and generation through a shared visual token space.

The key challenge is to construct visual tokens that retain the semantic information required for understanding while preserving sufficient visual detail for faithful reconstruction. We first analyze this tension and identify a propagation bottleneck in the vision encoder hierarchy (Sec. 2.1). Based on this analysis, we introduce Abstraction Fusion (AF), which aggregates representations across different levels of the pretrained vision encoder to form unified visual tokens for both understanding and generation (Sec. 2.2). Finally, we introduce Block-wise MoT, which integrates a lightweight generative stream into the pretrained VLM at block boundaries (Sec. 2.3).

2.1 Propagation Bottleneck

Our goal is to extend the representations learned by a pretrained vision encoder to support image reconstruction while preserving their semantic quality. A key challenge is that representations



(a) Overall tokenizer architecture (b) Semantics and reconstruction during training

Figure 2: **Abstraction Fusion.** (a) AF aggregates multi-level representations from a frozen vision encoder using per-layer normalization and projection, self-attention, and linear pooling, followed by a 256-dimensional VAE bottleneck and dual semantic and pixel decoders. (b) Semantic similarity and reconstruction quality over the first 20k tokenizer-training steps under the same VAE bottleneck and decoder architecture. Freezing encoders preserves semantics but limits reconstruction, while encoder fine-tuning exhibits a semantics–reconstruction trade-off. AF maintains near-frozen semantic similarity while achieving better reconstruction quality than the fine-tuned encoder.

become increasingly abstract across encoder depth, making fine-grained visual details difficult to recover from the final representation (Mahendran & Vedaldi, 2015; Li et al., 2024c). Adapting the vision encoder for reconstruction improves reconstruction fidelity but can degrade semantic quality, reflecting the semantics–reconstruction trade-off shown in Fig. 1c, consistent with observations in prior unified representation learning (Wu et al., 2025c; Ma et al., 2025; Yang et al., 2026).

Prior work has primarily attributed the semantics–reconstruction trade-off to optimization difficulties or limited representation capacity (Ma et al., 2025; Yang et al., 2026; Fan et al., 2025). We investigate a complementary explanation based on how information is transformed across the depth of the vision encoder. Vision encoders progressively transform local visual information into more abstract semantic representations (Yosinski et al., 2014; Mahendran & Vedaldi, 2015), while intermediate layers retain fine-grained visual details that are increasingly attenuated in deeper representations (Li et al., 2024c; Zhang et al., 2024). When reconstruction relies only on the final representation, training may encourage deeper layers to preserve information characteristic of earlier representations, potentially limiting the progression toward semantic abstraction. We refer to this constraint as a *propagation bottleneck*.

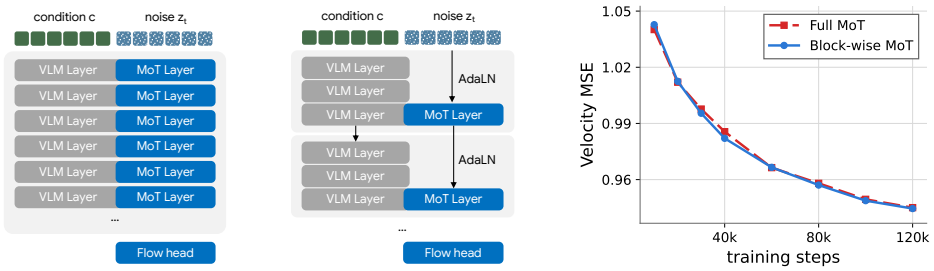
To examine this hypothesis, we analyze how joint training alters the representation hierarchy of the vision encoder. This view anticipates that deeper representations would become more similar to earlier-layer representations under joint training. We therefore measure the CKA similarity between each encoder layer and the initial-layer representation. As shown in Fig. 1b, joint training substantially increases the similarity between deeper and early-layer representations relative to the pretrained encoder. Increasing semantic supervision reduces this similarity, consistent with greater representational differentiation across depth. Together with the trade-off observed in Fig. 1c, these results are consistent with the propagation bottleneck hypothesis.

These observations suggest avoiding exclusive reliance on the final encoder representation for reconstruction. As illustrated in Fig. 1a, we leverage representations from multiple abstraction levels of the encoder, directly incorporating information available at different depths. This motivates Abstraction Fusion (AF), which constructs visual tokens by integrating representations across the encoder hierarchy.

2.2 Visual Tokenization with Abstraction Fusion

Based on the above analysis, we introduce Abstraction Fusion (AF), which constructs reconstruction-capable visual tokens without modifying the pretrained encoder's representation hierarchy. Rather than forcing reconstruction-relevant information to propagate through the encoder depth, AF directly aggregates representations from multiple levels of abstraction.

Abstraction Fusion. Given an image x, we extract intermediate representations $\{h^{(\ell)}\}_{\ell=0}^{L}$ from the frozen vision encoder. AF maps this feature hierarchy into a fused representation f while preserving



(a) Full MoT (b) Block-wise MoT (c) Full MoT vs. Block-wise MoT

Figure 3: **Block-wise MoT.** (a) Full MoT pairs every VLM layer with a generation-specific MoT layer. (b) Block-wise MoT instead groups consecutive VLM layers and performs generative interaction only at block boundaries, with timestep conditioning through adaLN-single. (c) Under the same training setup, Block-wise MoT closely matches full MoT in timestep-aggregated validation velocity MSE during text-to-image training.

the spatial token structure:

$$f = \mathrm{AF}\Big(\{h^{(\ell)}\}_{\ell=0}^{L}\Big), \qquad h^{(\ell)} \in \mathbb{R}^{N\times D}, \quad f \in \mathbb{R}^{N\times D_f}, \tag{1}$$

where N denotes the number of spatial tokens and L the number of encoder layers.

We implement AF with three key components illustrated in Fig. 2a. Per-layer normalization and projection first map representations from different encoder layers into a common feature space. For each spatial position, we then aggregate information across abstraction levels by applying self-attention over the layer-wise features. A linear pooler reduces the resulting layer-wise features into a single fused token. The fused representation is finally compressed through a variational bottleneck into visual tokens $z \in \mathbb{R}^{N\times c}$, where c denotes the latent dimension.

Dual decoders. To shape the fused latent into a shared space for both understanding and reconstruction, we supervise z with two complementary decoders. A semantic decoder $\mathrm{Dec}_{\mathrm{sem}}$ reconstructs the pretrained VLM visual representation z_{sem}, while a pixel decoder $\mathrm{Dec}_{\mathrm{rec}}$ reconstructs the input image:

$$\hat{z}_{\mathrm{sem}} = \mathrm{Decsem}(z), \qquad \hat{x} = \mathrm{Decrec}(z). \tag{2}$$

We supervise the semantic decoder with $\mathcal{L}_{\mathrm{sem}} = \|\hat{z}_{\mathrm{sem}} - z_{\mathrm{sem}}\|_2^2 + r\big(1 - \cos(\hat{z}_{\mathrm{sem}}, z_{\mathrm{sem}})\big)$, and the pixel decoder with $\mathcal{L}_{\mathrm{rec}} = \|\hat{x} - x\|_1 + \mathrm{LPIPS}(\hat{x}, x)$. The overall tokenizer objective is

$$\mathcal{L}_{\mathrm{AF}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \beta\mathcal{L}_{\mathrm{KL}}, \tag{3}$$

where r weights the cosine term and $\mathcal{L}_{\mathrm{KL}}$ regularizes the variational posterior toward a standard Gaussian. We keep the pretrained vision encoder frozen and optimize only the AF module, variational bottleneck, semantic decoder, and pixel decoder. After joint training, we further fine-tune the pixel decoder $\mathrm{Dec}_{\mathrm{rec}}$ with an additional GAN loss.

We compare AF with frozen and fine-tuned encoder baselines while keeping the VAE bottleneck and dual-decoder architecture fixed. As shown in Fig. 2b, the frozen encoder preserves semantic similarity but provides limited reconstruction quality, while fine-tuning improves reconstruction at the cost of semantic similarity. AF substantially improves reconstruction while maintaining semantic similarity close to that of the frozen encoder.

2.3 Conditional Generation with Block-wise MoT

Given the AF visual tokens z, we next model conditional generation while keeping the pretrained VLM frozen. Flow matching requires processing noisy visual latents and timestep conditioning that are absent from VLM pretraining. We therefore introduce Block-wise MoT, a parameter-efficient variant of Mixture-of-Transformers (MoT; Liang et al. 2025) that couples a separate generative stream to the frozen VLM at selected depths.

Block-wise MoT. Standard MoT couples each VLM layer with a generation-specific expert, enabling interaction throughout the transformer hierarchy but introducing substantial parameter overhead (Fig. 3a). Block-wise MoT instead groups consecutive VLM layers into blocks and performs generative interaction only at block boundaries (Fig. 3b). This retains conditioning at multiple levels of the frozen VLM while substantially reducing the number of generation-specific parameters.

Generative stream. Given a noisy visual latent z_t, we initialize the generative stream as $g_0 = \mathrm{MLP}(z_t)$. Let $\ell_b{}_{b=1}^{B}$ denote the boundaries of B VLM blocks. The generative stream is updated once at each block boundary:

$$g_b = \mathcal{G}_b\left(g_{b-1}, h_\theta^{(\ell_b)}, t\right), \qquad b = 1, \ldots, B, \tag{4}$$

where $h^{(\ell_b)}(c)$ is the representation produced by the frozen VLM at boundary ℓ_b, and $\mathcal{G}_b$ is the corresponding generation block.

Each generation block uses the generative stream as queries and attends jointly to key–value representations from the generative stream and the corresponding frozen VLM features. This interaction is one-way: the generative stream receives VLM conditioning, while the frozen VLM representations remain unaffected by the noisy visual latents. We inject the flow timestep into every generation block using adaLN-single (Chen et al., 2023).

Flow prediction. After the final generation block, g_B provides multimodal conditioning for a flow prediction head, which estimates the velocity field from the noisy latent z_t and timestep t:

$$v_\phi(z_t, t, c) = \mathrm{Pred}_\phi(z_t, t, g_B), \tag{5}$$

where g_B is the final generative representation conditioned on the multimodal input c through the frozen VLM.

Given $z_t = (1-t)z + t\epsilon$, where $t \sim \mathcal{U}(0,1)$ and $\epsilon \sim \mathcal{N}(0, I)$, we optimize

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\left\|v_\phi(z_t, t, c) - (\epsilon - z)\right\|_2^2\right]. \tag{6}$$

For the prediction head, we adopt the wide-head DDT architecture (Wang et al., 2026; Zheng et al., 2026). As shown in Fig. 3c, Block-wise MoT closely matches the flow-matching loss of full MoT throughout training while using substantially fewer generation-specific parameters.

3 Experiments

We evaluate *Integral* to examine whether a pretrained vision-language models can be extended with strong visual generation while preserving its understanding capability. Specifically, we study three questions:

- **RQ1.** Can *Integral* add strong visual generation while preserving its multimodal understanding knowledge? (Tab. 1)
- **RQ2.** Can *Integral* leverage the its multimodal knowledge across diverse generation tasks? (Tab. 2, Tab. 4, Tab. 3)
- **RQ3.** How does the unified visual token space bridge understanding and generation in a pretrained VLM? (Fig. 4, Tab. 6, Tab. 5)

3.1 Setup

Model. We instantiate *Integral* with Qwen3.5-9B (Qwen Team, 2026) and keep the pretrained VLM frozen throughout training. We set the AF latent dimension to $c = 256$ and use eight generation blocks in the Block-wise MoT. The Block-wise MoT contains 0.89B parameters.[1]

Training data. We train the AF tokenizer on DataComp-1B (Gadre et al., 2023), BLIP-3o (Chen et al., 2025a), and RenderedText (Wendler, 2024). For generation training, we use BLIP-3o,

[1]Following the convention used for the compared models, we exclude visual tokenizer parameters from model-size comparisons.

Table 1: **Multimodal understanding and generation.** † a score obtained with rewritten prompts, and § a score obtained with chain-of-thought reasoning. ‡ Emu3-Chat and Emu3-Gen are reported in one row for convenience.

		Visual Understanding									Image Generation			
Model	Params	POPE	MME-P	MME-C	MMB	SEED	SEED2+	MMStar	MMMU	OCRBench	GenEval	GenEval2	GenEval++	DPG
Understanding only														
LLaVA-v1.5 (Liu et al., 2023)	7B	85.9	1511	–	64.3	65.8	–	33.1	35.7	31.8	–	–	–	–
LLaVA-OV (Li et al., 2024a)	7B	–	1580	418	80.8	75.4	62.2	61.7	48.8	62.2	–	–	–	–
Qwen2.5-VL (Bai et al., 2025b)	7B	–	–	–	83.5	–	70.5	–	58.6	83.7	–	–	–	–
GPT-5.2 (OpenAI, 2025a)	–	–	–	–	88.2	–	–	77.1	86.7	80.7	–	–	–	–
Gemini-3 Pro (Google DeepMind, 2025a)	–	–	–	–	93.7	–	–	83.1	87.2	90.4	–	–	–	–
Qwen3.5-9B (Qwen Team, 2026)	9B	89.6	1716	685	86.3	78.9	73.1	75.9	73.2	88.1	–	–	–	–
Generation only														
SD3-Medium (Esser et al., 2024)	7.7B	–	–	–	–	–	–	–	–	–	0.74	–	0.296	84.08
FLUX.1-dev† (Labs, 2024)	17B	–	–	–	–	–	–	–	–	–	0.82	–	0.314	83.84
FLUX-Kontext (Labs, 2025b)	–	–	–	–	–	–	–	–	–	–	–	–	0.343	–
HiDream-I1-Full (Cai et al., 2025)	31B	–	–	–	–	–	–	–	–	–	0.83	–	–	85.89
GPT-4o-Image (OpenAI, 2024)	–	–	–	–	–	–	–	–	–	–	0.84	–	0.739	85.15
Seedream 3.0 (Gao et al., 2025)	–	–	–	–	–	–	–	–	–	–	0.84	–	–	88.27
Qwen-Image (Wu et al., 2025a)	29B	–	–	–	–	–	–	–	–	–	0.87	33.8	–	88.32
Unified understanding and generation														
Emu3‡ (Wang et al., 2024)	8.5B	85.2	1244	–	58.5	68.2	44.6	–	31.6	68.7	0.54	–	–	80.60
Janus (Wu et al., 2024)	1.7B	87.0	1338	222	69.4	63.7	–	–	30.5	–	0.61	–	–	79.68
Transfusion (Zhou et al., 2025)	7.3B	–	–	–	–	–	–	–	–	–	0.63	–	–	–
Show-o (Xie et al., 2025a)	1.4B	80.0	1097	248	–	51.5	–	–	27.4	–	0.68	–	–	–
Show-o2 (Xie et al., 2025b)	8.9B	–	1621	–	–	69.8	61.3	56.6	48.9	32.4	0.76	–	–	**86.14**
Harmon (Wu et al., 2025b)	2.7B	87.6	1155	321	65.5	67.1	41.6	35.3	38.9	11.2	0.79	–	–	–
Janus-Pro (Chen et al., 2025b)	7B	87.4	1567	260	79.2	72.1	56.3	48.3	41.0	59.0	0.80	–	0.246	84.19
BAGEL (Deng et al., 2025)	14B	–	1687	**701**	85.0	78.5	71.9	–	55.3	73.3	0.82	23.1§	0.371	–
Ming-UniVision (Huang et al., 2025)	18B	–	–	–	–	–	56.8	–	40.3	72.4	0.85	–	–	82.12
***Integral* (ours)**	9.8B	**89.6**	**1716**	695	**85.4**	**78.8**	**73.2**	**76.0**	**73.5**	**87.3**	**0.88**	**39.68**§	**0.704**	84.41

ScaleEdit-12M (Chen et al., 2026), and RenderedText (Wendler, 2024), followed by supervised fine-tuning on the data mixture described in Appendix A.

Training. We first train the AF tokenizer and then train the generative pathway in three stages: low-resolution pretraining, higher-resolution interleaved training, and supervised fine-tuning. Full training schedules and hyperparameters are provided in Appendix A.

Evaluation. We evaluate *Integral* on visual understanding, text-to-image generation, and image editing. For visual understanding, we evaluate on POPE (Li et al., 2023b), MME (Fu et al., 2023), MMBench (Liu et al., 2024a), SEED-Bench (Li et al., 2023a), SEED-Bench-2-Plus (Li et al., 2024b), MMStar (Chen et al., 2024), MMMU (Yue et al., 2024), and OCRBench (Liu et al., 2024b). For text-to-image generation, we evaluate compositional and instruction-following capabilities on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), GenEval2 (Kamath et al., 2025), and GenEval++ (Ye et al., 2025a), and knowledge-intensive generation on WISE^{V}, the verified version of WISE (Niu et al., 2026). We evaluate image editing on ImgEdit-Bench (Ye et al., 2025b) and KRIS-Bench (Wu et al., 2025d) to assess multimodal conditional generation.

3.2 Main Results

Integral combines strong visual understanding and generation within a single model (Tab. 1).

Visual understanding. *Integral* preserves the strong understanding capability of the pretrained VLM. Among unified models in Tab. 1, *Integral* achieves strong understanding performance across benchmarks, outperforming larger UMMs such as BAGEL (14B) and Ming-UniVision (18B) on most reported metrics. Notably, BAGEL and Ming-UniVision include substantial generation-specific capacity. In contrast, *Integral* shares nearly all of its model capacity between understanding and generation.

Image generation. With only 0.89B generation-specific parameters, *Integral* achieves strong image generation performance across compositional benchmarks. As shown in Tab. 1, *Integral* reaches 0.88 on GenEval, surpassing larger UMMs such as BAGEL (14B) and Ming-UniVision (18B), as well as the generation-specialized Qwen-Image (29B). *Integral* remains strong on more challenging compositional benchmarks, achieving 39.68 on GenEval2 with chain-of-thought reasoning, compared with 23.1 for BAGEL under the same setting. It also substantially outperforms existing UMMs on GenEval++. These results suggest that *Integral* effectively draws on its multimodal understanding capabilities for generation.

Knowledge-intensive image generation. *Integral* effectively leverages its world knowledge for visual generation, achieving a WISE^{V} (Niu et al., 2026) score of 0.558 without chain-of-thought

Table 2: **Knowledge-intensive text-to-image generation on WISE^{V}.** Rows marked + *CoT* explicitly reason before generation. Bold marks the best overall score among the unified models.

Model	Params	Culture	Time	Space	Biology	Physics	Chemistry	Overall↑
Closed-source								
NanoBanana-Pro (Google DeepMind, 2025b)	–	0.8975	0.8167	0.9333	0.8167	0.8667	0.8750	0.8760
GPT Image 1.5 (OpenAI, 2025b)	–	0.8900	0.6917	0.8833	0.8000	0.7583	0.7750	0.8250
Qwen-Image-2.0 (Zhao et al., 2026)	–	0.8219	0.6500	0.8992	0.7917	0.8000	0.7479	0.7954
Open-source generation-only								
SD-3.5-large (Stability AI, 2024)	14B	0.4900	0.4083	0.4417	0.3000	0.3750	0.2083	0.4040
Qwen-Image (Wu et al., 2025a)	29B	0.6275	0.5250	0.5583	0.3417	0.4833	0.2500	0.5100
HunyuanImage-3.0 (Cao et al., 2025)	83B	0.5250	0.3917	0.4833	0.3083	0.4500	0.2417	0.4350
FLUX.2-dev (Labs, 2025a)	56B	0.6650	0.5667	0.6583	0.3667	0.5250	0.3750	0.5650
Unified understanding and generation								
Janus-Pro-7B (Chen et al., 2025b)	7B	0.3700	0.3500	0.2833	0.2833	0.4000	0.2333	0.3340
BAGEL (Deng et al., 2025)	14B	0.4125	0.3500	0.3083	0.2000	0.4417	0.2583	0.3520
***Integral* (ours)**	9.8B	0.6475	0.5500	0.5583	0.4667	0.5500	0.3667	**0.5580**
Chain-of-thought reasoning								
BAGEL + CoT	14B	0.7800	0.6333	0.5667	0.3750	0.5500	0.5083	0.6280
***Integral* (ours) + CoT**	9.8B	0.8700	0.6972	0.7806	0.7000	0.7056	0.6778	**0.7753**

Table 4: **Instruction-based image editing on ImgEdit-Bench.**

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall↑
GPT-4o (OpenAI, 2024)	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Instruct-Pix2Pix (Brooks et al., 2022)	2.45	1.83	1.41	2.01	1.44	1.44	3.55	1.20	1.46	1.88
MagicBrush (Zhang et al., 2023)	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
AnyEdit (Yu et al., 2024)	3.18	2.95	1.14	2.49	2.21	2.88	3.82	1.56	2.65	2.45
UltraEdit (Zhao et al., 2024)	3.44	2.81	2.00	2.96	2.45	2.83	3.76	1.91	2.98	2.70
ICEdit (Zhang et al., 2025b)	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit (Liu et al., 2025)	3.88	3.41	1.76	3.40	2.83	3.16	3.63	2.52	2.52	3.06
BAGEL (Deng et al., 2025)	3.56	3.31	1.88	2.62	2.88	3.44	4.49	2.38	4.17	3.20
OmniGen2 (Wu et al., 2026a)	3.74	3.54	1.77	3.21	2.77	3.57	4.81	2.30	4.14	3.43
FLUX-Kontext (Labs, 2025b)	3.80	3.69	1.81	4.20	2.97	3.89	4.62	3.00	4.18	3.60
***Integral* (ours)**	4.55	3.47	3.84	2.95	3.98	4.15	2.77	3.26	3.87	**3.65**

prompting. It substantially outperforms BAGEL (0.352) and closely approaches the much larger FLUX.2-dev (0.565, 56B). With chain-of-thought reasoning, *Integral* further improves to 0.775, compared with 0.628 for BAGEL under the same setting. These results provide further evidence that *Integral*'s world knowledge can be effectively utilized for visual generation.

Image editing. We show that *Integral* effectively leverages its multimodal knowledge not only for text-conditioned generation, but also for multimodal conditional generation. On ImgEdit-Bench (Ye et al., 2025b), *Integral* achieves a score of 3.65, outperforming unified models such as BAGEL as well as the editing-specialized FLUX.1 Kontext (Labs, 2025b) (3.60). *Integral* also performs strongly on knowledge-intensive image editing, achieving 52.72 overall on KRIS-Bench and outperforming BAGEL (47.76) as well as the specialized editing models in Tab. 3. Together, these results suggest that *Integral*'s multimodal understanding and world knowledge can support generation tasks that require jointly interpreting visual and textual inputs.

Table 3: **Knowledge-intensive image editing on KRIS-Bench.**

Model	Fact.	Concept.	Proc.	Overall↑
GPT-4o (OpenAI, 2024)	79.80	81.37	78.32	80.09
Gemini 2.0	65.26	59.65	62.90	62.41
InsPix2Pix (Brooks et al., 2022)	23.33	25.59	17.28	22.82
OmniGen (Xiao et al., 2025)	33.11	28.02	23.89	28.85
MagicBrush (Zhang et al., 2023)	41.84	39.24	26.54	37.15
AnyEdit (Yu et al., 2024)	39.26	41.88	31.74	38.55
Emu2	45.40	37.54	34.91	39.70
Step1X-Edit (Liu et al., 2025)	45.52	48.01	31.82	43.29
BAGEL (Deng et al., 2025)	47.71	52.17	40.23	47.76
***Integral* (ours)**	56.90	53.97	44.87	**52.72**

Taken together, these results show that *Integral* achieves strong understanding and generation by sharing most of its model capacity across both capabilities, with only 0.89B generation-specific parameters. Fig. 5 further illustrates these results with qualitative examples of image generation and editing.

3.3 Analysis

We next analyze how the unified visual token space bridges understanding and generation.

Table 5: **Visual understanding preservation.** Performance before and after applying AF; Avg. Retention$_{20}$ averages retention over 20 benchmarks provided in Appendix B.1.

Model	POPE↑	MME↑	MMB↑	SEED↑	SEED2+↑	MMStar↑	MMMU↑	RWQA↑	OCRBench↑	Avg. Retention$_{20}$ ↑
Qwen3.5-9B	89.6	2400	86.3	78.9	73.1	75.9	73.2	78.2	88.1	100.0%
Qwen3.5-9B w/ AF	89.6	2412	85.4	78.8	73.2	76.0	73.5	78.7	87.3	99.8%

Table 6: **Image reconstruction quality.** Reconstruction on ImageNet-1k at 256×256 resolution.

Visual Tokenizer	rFID↓	PSNR↑	SSIM↑
Unified tokenizers			
RAE w/ Dinov2 (Zheng et al., 2026)	0.49	18.1	0.478
RAE w/ SigLIP2 (Zheng et al., 2026)	0.44	21.3	0.593
VTP (Yao et al., 2026)	0.36	25.8	0.741
PS-VAE (Zhang et al., 2025a)	0.20	28.8	0.817
Reconstruction-only tokenizers			
SD-VAE (Rombach et al., 2022)	0.90	24.2	–
VA-VAE (Yao et al., 2025)	0.28	28.0	0.790
FLUX.1-VAE (Labs, 2024)	0.17	31.1	–
Qwen3.5-9B w/ AF (ours)	**0.15**	**32.2**	**0.942**

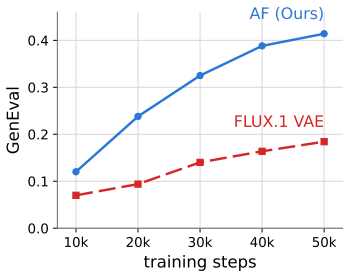


Figure 4: **Effect of the unified visual token space on generation.** GenEval during 50k-step text-to-image training.

AF preserves semantic information for visual understanding. Despite transforming the vision encoder representations into a unified visual token space, AF nearly preserves the original visual understanding capability. Across 20 benchmarks, Qwen3.5-9B with AF retains 99.8% of the base model's performance on average. The small deviations observed across benchmarks likely stem from the approximation error introduced by AF. Tab. 5 reports a representative subset, with the full results provided in Appendix B.1.

AF retains fine-grained visual information for high-fidelity reconstruction. AF preserves not only semantic information for understanding, but also the fine-grained visual information required for image reconstruction. As shown in Tab. 6, AF achieves an rFID of 0.15 and a PSNR of 32.2, outperforming prior unified tokenizers. Notably, its reconstruction quality is also competitive with dedicated reconstruction tokenizers, surpassing FLUX.1-VAE in both rFID and PSNR.

Sharing the visual token space improves generative learning. We compare AF with FLUX.1-VAE, which provides comparable reconstruction quality, while keeping the generative model and training setup unchanged. As shown in Fig. 4, AF consistently achieves higher GenEval scores, reaching 0.41 after 50k steps compared with 0.18 using FLUX.1-VAE latents. This gap suggests that reconstruction quality alone is insufficient; sharing a unified visual token space across understanding and generation is important for effective generative learning.

Overall, these results suggest that the unified visual token space effectively bridges understanding and generation, and that preserving the representation space used for understanding facilitates generative learning.

4 Related Work

Unified multimodal models. Unified multimodal models (UMMs) aim to integrate visual understanding and generation within a single framework. Early approaches pursued native unification by representing different modalities as discrete tokens and modeling them with a common autoregressive objective (Lu et al., 2024; Team, 2024; Wang et al., 2024). Recent work has increasingly incorporated continuous visual representations with diffusion- or flow-based generation objectives alongside autoregressive language modeling (Zhou et al., 2025; Xie et al., 2025b). To accommodate the resulting heterogeneity between understanding and generation, several approaches introduce task- or modality-specific pathways while retaining varying degrees of shared computation. For example, Janus decouples the visual encoding pathways for understanding and generation (Wu et al., 2024; Chen et al., 2025b), while some approaches introduce specialized transformer experts while retaining shared cross-modal interaction (Liang et al., 2025; Deng et al., 2025; Shi et al., 2025). Other approaches couple pretrained visual language models with dedicated generative path-



Figure 5: **Qualitative examples of generation and editing.** Generation uses diverse text prompts, while editing shows examples from ImgEdit (Ye et al., 2025b) and KRIS-Bench (Wu et al., 2025d).

ways (Pan et al., 2025; 2026). In contrast, *Integral* addresses this separation at the representation level by constructing a shared visual token space for both understanding and generation.

Visual representations for understanding and generation. Visual understanding and generation have traditionally relied on different types of visual representations. Semantic encoders learn transferable, high-level visual features, whereas generative tokenizers are primarily trained to retain information needed for image reconstruction (Radford et al., 2021; Oquab et al., 2023; Esser et al., 2021). Recent work has shown that pretrained semantic representations can also serve as effective latent spaces for generative modeling (Zheng et al., 2026). In parallel, unified visual representation learning has sought to combine semantic and reconstructive information within a shared representation (Ma et al., 2025; Yao et al., 2026; Zhang et al., 2025a; Yang et al., 2026; Fan et al., 2025). These approaches typically balance semantic preservation and reconstruction quality, and recent studies have analyzed the resulting trade-off from perspectives including representation capacity, optimization conflict, and structural decoupling (Ma et al., 2025; Yang et al., 2026; Yue et al., 2026). *Integral* instead builds on the pretrained representation hierarchy and aggregates information across multiple abstraction levels to construct unified visual tokens for both understanding and generation.

5 Conclusion

We introduced *Integral*, a unified multimodal model that extends a pretrained vision-language model with strong visual generation while preserving multimodal understanding. Despite using only 0.89B generation-specific parameters, *Integral* achieves strong performance across compositional generation, knowledge-intensive generation, and multimodal image editing. Our results show that sharing the visual representation space between understanding and generation is key to transferring multimodal knowledge to visual generation.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We did not use generative AI tools for other tasks requiring disclosure. Additionally, we used generative AI tools to create or edit software code and scientific figures, source and search for information, identify relevant literature, summarize or analyze existing literature, and edit the paper to improve readability. We reviewed and verified all AI-assisted work, including code, figures, referenced information, and edited text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Integral architecture and training objectives in Sec. 2, and the experimental setup and evaluation protocols in Sec. 3.1. Appendix A provides the complete training pipeline, datasets, and stage-wise training procedure, while Table 7 summarizes the optimization hyperparameters, training schedules, objectives, and data sampling ratios used in all stages. Additional implementation details and experimental results are provided in the appendix.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Guanzhou Chen, Erfei Cui, Changyao Tian, Danni Yang, Ganlin Yang, Yu Qiao, Hongsheng Li, Gen Luo, and Hongjie Zhang. Scaleedit-12m: Scaling open-source image editing data generation via multi-agent framework. *arXiv preprint arXiv:2603.20644*, 2026.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

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

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, Pengfei Wan, and Yi-Fan Zhang. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing, 2025c. URL https://arxiv.org/abs/2509.24900.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 28652–28668. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Clark_Molmo2_Open_Weights_and_Data_for_Vision-Language_Models_with_Video_CVPR_2026_paper.html.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. *arXiv preprint arXiv:2512.19693*, 2025.

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

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

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

Google DeepMind. Gemini 3 pro model card. Model card, 2025a. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Google DeepMind. Gemini 3 pro image external model card. Model card, 2025b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

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

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Jiwon Kang, Heeji Yoon, Jaewoo Jung, Jaewon Min, Minkyeong Jeon, Biyeon Hwang, Sangwon Jung, and Seungryong Kim. Transferability between understanding and generation in unified multimodal models. In *European Conference on Computer Vision (ECCV)*, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*, pp. 235–251. Springer, 2016. doi: 10.1007/978-3-319-46493-0_15.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025a.

Black Forest Labs. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. https://arxiv.org/abs/2506.15742, 2025b.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision-language reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c6XIVI3TiQ.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://openreview.net/forum?id=xozJw0kZXF.

Yunheng Li, ZhongYu Li, Quansheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. *arXiv preprint arXiv:2406.00670*, 2024c.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Nu6N69i8SB.

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

Yiheng Lin, Siyu Jiao, Xiaohan Lan, Wei Zhou, Qi She, Fei Yu, Heyun Chen, Zhengwei Wang, Jinghuan Chen, Moran Li, Yingchen Yu, Zijian Feng, Yao Zhao, Yunchao Wei, and Yujie Zhong. Textsculptor: Training and benchmarking scene text editing. *arXiv preprint arXiv:2605.21090*, 2026.

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

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xiangyue Liu, Zijian Zhang, Miles Yang, Zhao Zhong, Liefeng Bo, and Ping Tan. Symbiotic-moe: Unlocking the synergy between generation and understanding. *arXiv preprint arXiv:2604.07753*, 2026a.

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

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Zeyu Liu, Zanlin Ni, Yang Yue, Cheng Da, Huan Yang, Di Zhang, Kun Gai, and Gao Huang. Steering visual generation in unified multimodal models with understanding supervision. *arXiv preprint arXiv:2605.05781*, 2026b.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15740–15751, 2026c.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26429–26445. IEEE, 2024.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, BINGYUE PENG, and Xiaojuan Qi. Unitok: a unified tokenizer for visual generation and understanding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 129274–129297. Curran Associates, Inc., 2025. doi: 10.52202/085713-4306. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bbe04d329d531e8814f1199098bd8fb6-Paper-Conference.pdf.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

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

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

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. In *Forty-third International Conference on Machine Learning*, 2026.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. Update to GPT-5 system card: GPT-5.2. System card, 2025a. URL https://openai.com/index/gpt-5-system-card-update-gpt-5-2/.

OpenAI. The new ChatGPT images is here. OpenAI blog post, introducing GPT-Image-1.5, 2025b. URL https://openai.com/index/new-chatgpt-images-is-here/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Xichen Pan, Aashu Singh, Satya Narayan Shukla, Xiangjun Fan, Shlok Kumar Mishra, and Saining Xie. Repfusion: Leveraging multimodal priors for denoising in representation space. *arXiv preprint arXiv:2606.14700*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Lin, Luke Zettlemoyer, and LILI YU. Lmfusion: Adapting pretrained language models for multimodal generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 9163–9182. Curran Associates, Inc., 2025. doi: 10.52202/085713-0310. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0d33b1148c6ee6bb3ea9f056ae3301e6-Paper-Conference.pdf.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Stability AI. Stable diffusion 3.5 large. Model card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-large.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. doi: 10.48550/arXiv.2405.09818. URL https://github.com/facebookresearch/chameleon.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, et al. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 40633–40642, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, JingJing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025b. doi: 10.48550/ARXIV.2508.18265. URL https://doi.org/10.48550/arXiv.2508.18265.

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

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025c.

Chris Wendler. Renderedtext, 2024. URL https://huggingface.co/datasets/wendlerc/RenderedText.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21964–21975, June 2026a.

Penghao Wu, Haiwen Diao, Weichen Fan, Lewei Lu, Dahua Lin, and Ziwei Liu. Uncovering understanding-generation synergy in native unified multimodal models: From representation, task to system. *arXiv preprint arXiv:2609.01607*, 2026b.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17739–17750. IEEE, 2025b.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e9e140df6de01afb672cb859d203c307-Paper-Conference.pdf.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025d.

xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v, 2024. Introduces the RealWorldQA benchmark.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025a.

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

Panqi Yang, Haodong Jing, Jiahao Chao, Tingyan Xiang, Li Lin, Yao Hu, Yang Luo, and Yongqiang Ma. Muse: Resolving manifold misalignment in visual tokenization via topological orthogonality. *arXiv preprint arXiv:2605.05646*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. In *European Conference on Computer Vision*, pp. 661–678. Springer, 2026.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025a.

Sen Ye, Mengde Xu, Shuyang Gu, Di He, Liwei Wang, and Winston Hu. Understanding vs. Generation: Navigating Optimization Dilemma in Multimodal Models. In *International Conference on Learning Representations*, 2026. URL https://mlanthology.org/iclr/2026/ye2026iclr-understanding/.

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

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, Yi Wang, Limin Wang, and Yali Wang. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 84741–84772, 2026.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023.

Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian Ge, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, et al. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. *arXiv preprint arXiv:2512.17909*, 2025a.

Yongjian Zhang, Longguang Wang, Kunhong Li, Yun Wang, and Yulan Guo. Learning representations from foundation models for domain generalized stereo matching. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2024.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b. arXiv:2504.20690.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale, 2024. URL https://arxiv.org/abs/2407.05282.

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

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

A TRAINING DETAILS

We train *Integral* in two phases: visual tokenizer training and generation training. Visual tokenizer training consists of three stages: (i) initial 256×256 training, (ii) dynamic-resolution training up to 512×512, and (iii) GAN decoder training. Generation training also consists of three stages: (i) text-to-image pretraining, (ii) interleaved pretraining, and (iii) interleaved supervised fine-tuning (SFT). Table 7 summarizes the optimization settings, schedules, objectives, and data composition for all stages.

A.1 DATASETS

For visual tokenizer training, we use DataComp-1B (Gadre et al., 2023) for the initial 256×256 training. For dynamic-resolution training, we additionally use BLIP-3o (Chen et al., 2025a) and RenderedText (Wendler, 2024). For GAN decoder training, we use DataComp-1B.

For generation training, we use BLIP-3o, RenderedText, and TextAtlas5M (Wang et al., 2025a) for text-to-image pretraining. For interleaved pretraining, we use ScaleEdit-12M (Chen et al., 2026), BLIP-3o, TextSculpt-Data (Lin et al., 2026), RenderedText, TextAtlas5M, Pico-Banana-400K, and image-reconstruction data from ImageNet-1K (Russakovsky et al., 2015) and BLIP-3o. For interleaved SFT, we use GPT-Image-Edit-1.5M (Wang et al., 2025c), Echo-4o-Image (Ye et al., 2025a), OpenGPT-4o-Image (Chen et al., 2025c), ScaleEdit-12M, Nano-Consistent-150K, TextSculpt-Data, Zebra-CoT (Li et al., 2026), and image-reconstruction data from ImageNet-1K and BLIP-3o. We do not use BLIP3o-60K for interleaved SFT, as its instructions overlap with those in GenEval.

Table 7: **Training recipe for *Integral*.** Optimization settings, schedules, objectives, and high-level data composition for the three tokenizer-training stages and three generation-training stages.

	Tokenizer			Generation		
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
Optimization						
Learning rate	10^{-4}	$10^{-4} \rightarrow 10^{-5}$	2×10^{-5}	10^{-4}	10^{-4}	2×10^{-5}
LR scheduler	Constant	Cosine	Constant	Constant	Constant	Constant
Warm-up steps	0	500	0	0	1,000	500
Weight decay	0.01	0.01	0.0	0.01	0.01	0.01
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.95)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Gradient norm clip	1.0	1.0	1.0	1.0	1.0	1.0
EMA decay	–	–	–	0.9999	0.9999	0.999
Precision	bf16 autocast, FSDP2 parameters and gradients in fp32					
Schedule						
Training steps	200K	400K	50K	200K	50K	100K
GPUs (H100)	8	8	8	32	48	48–64
Batch per step	512	512	256	1024	393,216 tok	393,216 tok
Resolution	256	256–512	256	256	≤ 512	≤ 512
Objective						
KL weight	10^{-6}	10^{-6}	–	–	–	–
Adversarial weight	–	–	0.5	–	–	–
Flow matching	–	–	–	✓	✓	✓
Timestep distribution	–	–	–	Logit-normal $(0, 1)$, time shift 4.0		
Caption dropout	–	–	–	0.1	0.1	0.1
Loss weight (text : image)	–	–	–	–	1.0 : 1.0	1.0 : 1.0
Data sampling ratio						
Text-to-image	–	–	–	1.00	0.400	0.350
Image editing	–	–	–	–	0.595	0.549
Multi-reference	–	–	–	–	–	0.051
Interleaved reasoning	–	–	–	–	–	0.049
Image reconstruction	–	–	–	–	0.005	0.001

Table 8: **Visual understanding preservation.**

Model	SEED	SEED2+	MME-P	MME-C	MMMU	POPE	MMB	MMStar	HallusionBench	AI2D	ScienceQA
Qwen3.5-9B	78.9	73.1	1716	685	73.2	89.6	86.3	75.9	72.6	87.7	94.4
Qwen3.5-9B w/ AF	78.8	73.2	1716	695	73.5	89.6	85.4	76.0	72.5	87.8	94.6
Qwen3.5-2B	75.8	65.4	1585	554	57.0	89.0	74.4	68.9	66.7	81.3	82.4
Qwen3.5-2B w/ AF	75.7	65.0	1594	532	55.0	88.9	73.5	66.8	66.9	80.9	81.7

Model	RWQA	OCRBench	TextVQA	DocVQA	InfoVQA	MMVP	BLINK	VizWiz	CV-Bench^{2D}	CV-Bench^{3D}	Avg. Retention$_{20}$
Qwen3.5-9B	78.2	88.1	84.3	95.6	86.9	67.3	69.3	40.0	83.0	91.1	100.0%
Qwen3.5-9B w/ AF	78.7	87.3	83.9	95.6	86.6	67.3	67.9	39.9	83.3	90.6	99.8%
Qwen3.5-2B	67.7	83.2	79.1	93.5	79.3	46.7	23.2	37.7	71.6	81.9	100.0%
Qwen3.5-2B w/ AF	68.8	80.8	78.9	92.0	77.3	47.3	23.4	37.1	71.9	81.3	99.2%

A.2 Training

Visual tokenizer training. We first train the AF tokenizer at 256×256 resolution to learn a shared visual token space that preserves semantic information while supporting image reconstruction. We then continue with dynamic-resolution training up to 512×512 to extend the tokenizer to higher resolutions and diverse aspect ratios. Finally, we train a GAN-enhanced pixel decoder to improve perceptual reconstruction quality while keeping the learned visual token space fixed.

Generation training. We first perform text-to-image pretraining at 256×256 resolution to learn visual generation from text within the unified token space. We then continue with interleaved pretraining at resolutions up to 512×512, extending the model from text-conditioned generation to multimodal conditional generation with image inputs. Finally, we perform interleaved supervised fine-tuning on a broader mixture of generation tasks, including text-to-image generation, image editing, multi-reference generation, and interleaved reasoning.

B Additional Evaluation of Abstraction Fusion

We provide additional evaluations of Abstraction Fusion (AF) beyond the main experiments. We first examine whether AF preserves the pretrained understanding capability across a broader set of benchmarks, and then evaluate its generality across different model scales and vision encoders.

B.1 Understanding Preservation

We evaluate the understanding capability of *Integral* on an extended set of multimodal benchmarks.

Evaluation. Beyond the benchmarks reported in the main text, we additionally evaluate on HallusionBench (Guan et al., 2024), AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), RealWorldQA (xAI, 2024), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), MMVP (Tong et al., 2024b), BLINK (Fu et al., 2024), VizWiz (Gurari et al., 2018), and CV-Bench (Tong et al., 2024a).

Results. As shown in Table 8, *Integral* closely preserves the performance of Qwen3.5-9B across the evaluated tasks, retaining 99.8% of its average understanding performance. These results show that AF introduces negligible degradation to the pretrained understanding capability while enabling the same visual tokens to support image reconstruction.

B.2 Generality across Model Scales and Vision Encoders

We further examine whether AF generalizes beyond the Qwen3.5-9B configuration used in the main experiments. We apply the same AF design to Qwen3.5-2B and to SigLIP2 (`google/siglip2-so400m-patch16-256`). For the Qwen3.5 models, we measure semantic preservation by the retention of the original multimodal understanding performance, while for SigLIP2 we use ImageNet zero-shot classification accuracy.

As shown in Table 9, AF maintains both reconstructive and semantic information across different model scales and vision encoders. The Qwen3.5-2B variant achieves 32.33 dB PSNR while closely retaining the understanding capability of its base model. When applied to SigLIP2, AF achieves 34.08 dB PSNR while maintaining a zero-shot classification accuracy of 0.83. These results show

Table 9: **Generality of Abstraction Fusion across model scales and vision encoders.** Reconstruction is evaluated on ImageNet at 256×256. For Qwen3.5, we report understanding retention relative to the corresponding base model. For SigLIP2, we report ImageNet zero-shot classification accuracy.

	Reconstruction	Semantic Preservation	
Model	PSNR ↑	Understanding Retention ↑	Zero-shot Acc. ↑
Qwen3.5-9B (Qwen Team, 2026)	–	100.0%	–
+ AF	32.20	99.8%	–
Qwen3.5-2B (Qwen Team, 2026)	–	100.0%	–
+ AF	32.33	99.2%	–
SigLIP2 (Tschannen et al., 2025)	–	–	0.83
+ AF	34.08	–	0.83

that AF is not specific to the Qwen3.5-9B configuration and extends to both smaller models and different pretrained vision encoders.