OPENVISION 3: A FAMILY OF UNIFIED VISUAL ENCODER FOR BOTH UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

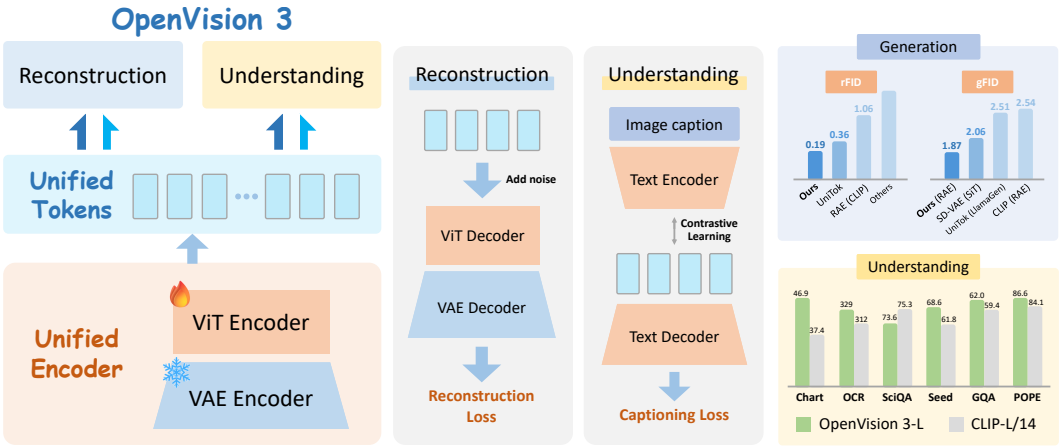


Figure 1: **An overview of OpenVision 3's architecture design and performance highlight**. **Left panel**: The architecture of OpenVision 3. We employ a frozen VAE and a trainable ViT as the unified tokenizer, which produces tokens that are fed simultaneously into both the reconstruction and understanding branches. **Middle panel**: The learning objectives of the two branches. For the reconstruction branch, we focus on pixel-level image reconstruction; concurrently, the understanding branch is optimized via joint contrastive and captioning objectives. **Right panel**: The performance summarization. The result shows that OpenVision 3 outperforms other unified tokenizers and semantics-based encoders in reconstruction and generation, while showing superior understanding ability when compared with CLIP.

ABSTRACT

This paper presents a family of advanced vision encoders, named OpenVision 3, that learns a *single*, *unified* visual representation that can serve both image understanding and image generation. Our core architecture is simple: we feed VAE-compressed image latents to a ViT encoder and train its output to support two complementary roles. First, the encoder output is passed to the ViT-VAE decoder to reconstruct the original image, encouraging the representation to capture generative structure. Second, the same representation is optimized with contrastive learning and image-captioning objectives, strengthening semantic features. By jointly optimizing reconstruction- and semantics-driven signals in a shared latent space, the encoder learns representations that *synergize* and *generalize* well across both regimes. We validate this unified design through extensive downstream evaluations with the encoder ***frozen***. For generation, we test it under the RAE framework: ours substantially surpasses the standard CLIP-based encoder (*e.g.*, gFID: 1.87 *vs*. 2.54 on ImageNet). For multimodal understanding, we plug the encoder into the LLaVA-1.5 and LLaVA-NeXT framework: it achieves performance on par with or superior to a standard CLIP vision encoder (*e.g.*, 59.2 *vs*. 58.1 on GQA and 35.0 *vs*. 32.0 on ChartQA), and exhibits remarkable scalability. We provide empirical evidence that generation and understanding are mutually beneficial

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.

in our architecture, while further underscoring the critical role of the VAE latent space. We hope this work can spur future research on unified modeling.

1 INTRODUCTION

Unified Multimodal Models (UMMs) have emerged as a cornerstone of multimodal research, driven by the need for systems that seamlessly integrate visual understanding and generation. Their development is grounded in the Platonic Representation Hypothesis (Huh et al., 2024), which posits that different data modalities reflect a shared underlying reality, and that learning a unified multimodal representation enables mutual benefits across modalities while improving generalization. The success of representative proprietary UMMs such as GPT-4o (Hurst et al., 2024) and Gemini-2.5 Flash (Comanici et al., 2025), as well as public models like BAGEL (Deng et al., 2025), further supports this view, showcasing strong capabilities in dialogue-based multi-turn generation, multimodal in-context learning, and fine-grained control over generated content.

A key challenge in developing native UMMs lies in how visual representations are encoded. Owing to the representational discrepancy between visual understanding and visual generation, a common UMM design, exemplified by UniFluid (Fan et al., 2025), BAGEL (Deng et al., 2025), and MOGAO (Liao et al., 2025), employs two distinct visual tokenizers that encode the same image twice, producing one set of high-level semantic tokens and another set of low-level, pixel-reconstructable tokens. While effective, this approach increases system complexity and may hinder deeper synergy between understanding and generation.

Another line of work attempts to bridge this gap through shared visual tokenizers. However, these approaches typically rely on quantized hidden representations, which inevitably introduce discretization errors and limit generation quality (*e.g.*, TokenFlow (Qu et al., 2025), UniTok (Ma et al., 2025), and EMU3.5 (Cui et al., 2025)). As a result, developing a simple yet effective *continuous* visual tokenizer that naturally supports both visual understanding and generation remains an open and practical challenge.

This paper presents OpenVision 3 as a step toward mitigating this challenge. Concretely, we build our tokenizer by stacking a ViT encoder on top of a well-trained VAE encoder. The output of the ViT encoder is further fed into two separate branches, one generation decoder that is trained to reconstruct the original image and enforce preservation of low-level visual information, and another understanding decoder that is trained by contrastive and captioning objectives, enhancing semantic supervision. Intriguingly, as analyzed in Section 5.1, this design choice can non-trivially *synergize* the learning of both fine-grained details and high-level semantics: adding reconstruction supervision modestly improves downstream understanding, while adding semantic supervision substantially improves downstream generation despite not improving direct reconstruction metrics. This behavior is also consistent with recent evidence that semantically informed tokenization can facilitate downstream generative modeling (Yu et al., 2024; Leng et al., 2025; Yao et al., 2025), and may serve as a direct drop-in replacement for purely reconstruction-oriented tokenizers (Zheng et al., 2025).

Our experiments validate the effectiveness of OpenVision 3 across understanding, reconstruction and generation. Crucially, in all downstream evaluations we keep the tokenizer/encoder **frozen**, ensuring that the reported gains reflect the quality and transferability of the learned visual representation rather than task-specific fine-tuning. For understanding evaluation, we integrate our tokenizer into the LLaVA-1.5 and LLaVA-NeXT framework for training and evaluate its performance across various standard multimodal benchmarks. For the reconstruction evaluation, we evaluate the quality of reconstructed images from OpenVision 3 on COCO (Lin et al., 2014) and ImageNet (Deng et al., 2009). For the generation evaluation, we train flow matching model following RAE (Zheng et al., 2025) on ImageNet. The results demonstrate that OpenVision 3 shows comparable or superior performance to CLIP in terms of understanding capabilities (*e.g.*, 59.2 *vs*. 58.1 on GQA and 35.0 *vs*. 32.0 on ChartQA), while surpassing existing unified tokenizers in image reconstruction (*e.g.*, rFID: 0.187 *vs*. 0.362 on ImageNet). For image generation, our tokenizer outperforms standard CLIP-based encoder under the RAE framework by a large margin (*e.g.*, gFID: 1.87 *vs*. 2.54 on ImageNet). Moreover, through the ablation study, we observed that our tokenizer achieves a mutual promotion between understanding and generation during training. The utilization of the VAE latent space has also been empirically validated as both effective and indispensable through our ex-

periments. We hope that releasing OpenVision 3 will catalyze further research into more advanced unified vision tokenizers.

2 Related Work

2.1 Vision-Language Pretraining

Vision-Language pretraining serves as the cornerstone of multimodal representation learning. Pioneering works, exemplified by CLIP, adopt contrastive learning as their core methodology to extract and align visual and textual features. This training paradigm was subsequently adopted by a wide range of studies, such as LAION (Schuhmann et al., 2022), DataComp (Gadre et al., 2023), DFN (Fang et al., 2023), OpenCLIP (Cherti et al., 2023), MetaCLIP (Xu et al., 2023; Chuang et al., 2025) and CLIPA (Li et al., 2023b;a). These studies focus primarily on efficient and open-sourced data and scaling methodologies. Follow-up works have continuously explored alternative training regimes. CoCa (Yu et al., 2022) adds a captioning loss on the multimodal decoder outputs which predicts text tokens autoregressively. SigLip (Zhai et al., 2023) proposes to replace contrastive loss with pairwise Sigmoid loss. SigLip2 (Tschannen et al., 2025) further extends this by incorporating captioning-based pretraining, self-distillation and masked prediction. The AM-RADIO (Ranzinger et al., 2024; Heinrich et al., 2025) series of works is dedicated to multi-resolution training and knowledge distillation from multiple teacher models. More recently, CLIPS (Liu et al., 2024b), OpenVision (Li et al., 2025), and OpenVision 2 (Liu et al., 2025a) have focused on the efficient utilization of captioning loss in vision-language pretraining, demonstrating it to be a low-cost yet high-performance approach. Our work builds upon this line of research and extends this efficient paradigm to unified multimodal learning. By combining contrastive, captioning, and reconstruction losses, we simultaneously supervise semantic and generative learning, resulting in reciprocal performance gains for both parts.

2.2 Unified Tokenizer

Extracting representative features for both generation and understanding has been a bottleneck for the development of unified modeling. Previous works mostly adopt separate encoders for the two kinds of features, and then combine them. For example, BAGEL takes FLUX-VAE (Labs, 2024; Labs et al., 2025) for low-level features and SigLIP2 for semantic features. UniWorld-V1 (Lin et al., 2025) also computes the two types of features separately and then concatenates them.

In contrast to above works, another line of studies focuses more on developing unified tokenizers that fuse semantic and pixel-level features. Inspired by the success of VQGAN (Esser et al., 2021), early unified tokenizers predominantly adapt a discrete token design. Discrete tokenizers rely on vector quantization (VQ) to train representative unified codebooks. For example, TokenFlow (Qu et al., 2025) jointly optimizes semantic and pixel-level features by incorporating dual codebooks with a shared mapping. VILA-U (Wu et al., 2024) discretizes the features extracted by the SigLIP with residual quantization. UniTok (Ma et al., 2025) uses multi-codebook quantization to construct unified discrete representations. Lately, the prevalent trend has gradually shifted toward continuous tokenizers. UniLIP (Tang et al., 2025) tailors CLIP features for image generation by incorporating self-distillation constraints to bridge the domain gap. Show-o2 (Xie et al., 2025) applies semantic and low-level projection to VAE latents and fuses dual features to produce unified feature space. More recently, the concurrent work TUNA (Liu et al., 2025b) further simplifies this direction by connecting a VAE and a ViT as a unified tokenizer, which is most related to our work. However, TUNA relies on pretrained ViT checkpoints and its tokenizer training recipe remains less transparent. Moreover, TUNA uses ViT tokens as conditions for a flow head to predict VAE latent velocity, and image generation is eventually carried out in the denoised VAE latent space. In contrast, our tokenizer directly learns image representations as the generation space itself. In image generation with our tokenizer, the generator directly generates the representations and the tokenizer decodes the representation to the final image. We train the ViT from scratch and propose an effective training paradigm for unified representations that support both understanding and generation.

3 METHOD

3.1 MOTIVATION

Developing a unified tokenizer is a pivotal step toward unifying generation and understanding, but it is often hindered by the difficulty of establishing a unified feature space and highly efficient training. Previous studies have presented impressive methods to eliminate these obstacles. However, explorations into constructing unified representations remain in their preliminary stages, and the associated training pipelines still remain non-transparent to the community. In the following, we present our model, which constructs unified vision representation space through a VAE and a ViT in an effective and straightforward way. We demonstrate to the research community how to train a unified tokenizer efficiently from scratch within the VAE latent space.

3.2 OPENVISION 3: A UNIFIED TOKENIZER

OpenVision 3 uses a VAE encoder and a vision transformer (ViT) to extract unified vision features, as depicted in Figure 1. The input image $x \in \mathbb{R}^{H\times W\times C}$ is first encoded by the VAE encoder $\mathcal{E}_{vae}$ from FLUX.1-dev into VAE latents z_{vae}, and the following training process is completely under the VAE latent space. Next, the VAE latents are fed into the ViT encoder $\mathcal{E}_{vit}$ to extract the unified representations z_u for both understanding tasks and generation tasks. During the VAE stage, the FLUX.1 VAE downsamples the image height and width by 8×, respectively. Therefore, we adjust the patch size of the ViT to 2×2 so that the whole compression ratio is 16×, which aligns with common settings. Formally,

$$z_{vae} = \mathcal{E}_{vae}(x) \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times D_{vae}} \tag{1}$$

$$z_u = \mathcal{E}_{vit}(z_{vae}) \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times D_u} \tag{2}$$

where D_{vae} is the VAE latent channels, D_u is the ViT dimensions. The encoded unified feature z_u then goes into the reconstruction branch and the understanding branch to do decoding. OpenVision 3 employs two distinct branches to cultivate its ability to extract both generative and interpretive vision representations. The two branches are completely separate, and their respective architectures will be elaborated upon below.

Reconstruction branch. The reconstruction decoding part mirrors the structure of the tokenizer, maintaining a near-symmetrical configuration. Before the decoding, we first add noise to the unified representations in order to improve the generalization of generation ability. The perturbed feature $\tilde{z}_u$ is generated by adding Gaussian noise scaled by a sample-specific intensity:

$$\tilde{z}_u = z_u + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{3}$$

where σ is uniformly sampled from $[0, \tau]$ for each instance in the batch, τ is a constant. Then we use a ViT decoder with patch size 1×1 and a linear layer to convert the noised unified feature $\tilde{z}_u$ back into VAE latents $\hat{z}_{vae}$. Next, the VAE decoder is applied to decode the $\hat{z}_{vae}$ into reconstruction image $\hat{x}$. The whole reconstruction loss includes the reconstruction loss of image $\hat{x}$ and VAE latents $\hat{z}_{vae}$, and a perceptual loss based on LPIPS. The whole reconstruction loss can be formulated as:

$$\mathcal{L}_{rec} = \ell_1(x, \hat{x}) + \beta\ell_1(z_{vae}, \hat{z}_{vae}) + \lambda\mathcal{L}_{LPIPS}(x, \hat{x}) \tag{4}$$

Understanding branch. The paradigm of understanding branch generally follows OpenVision (Li et al., 2025), where we do contrastive learning and image captioning. As shown in Figure 1, we use a text encoder to extract the caption feature z_{txt} to calculate contrastive loss with the unified visual feature z_u. In parallel, we utilize a text decoder to perform autoregressive prediction of synthetic captions from the unified representations and calculate the corresponding captioning loss. Formally, the understanding loss can be formulated as:

$$\mathcal{L}_{und} = \mathcal{L}_{caption} + \alpha\mathcal{L}_{contrastive}(z_u, z_{txt}) \tag{5}$$

The overall training objective is:

$$\mathcal{L}_{overall} = \omega_{rec}\mathcal{L}_{rec} + \omega_{und}\mathcal{L}_{und} \tag{6}$$

We configure ω_{und} as double that of ω_{rec} during the training process. Reducing ω_{rec} helps to preserve generative quality while ensuring that the understanding capability remains unimpaired.

Table 1: **Reconstruction performance of visual tokenizers.** Evaluations are performed on the ImageNet and COCO validation sets. Images are resized and center-cropped to 256×256.

Model	ImageNet				COCO			
	PSNR↑	SSIM↑	LPIPS↓	rFID↓	PSNR↑	SSIM↑	LPIPS↓	rFID↓
Generation-oriented Tokenizer								
SD-VAE	26.26	0.745	0.133	0.606	25.99	0.759	0.130	4.142
SD3-VAE	31.29	0.886	0.059	0.201	31.18	0.894	0.056	1.671
Cosmos	25.07	0.700	0.167	0.959	24.74	0.711	0.165	5.063
FLUX-VAE	32.86	0.917	0.044	0.176	32.73	0.923	0.041	1.343
Wan2.1-VAE	31.34	0.886	0.058	0.945	31.19	0.895	0.055	3.449
Unified Tokenizer								
RAE (CLIP)	17.44	0.403	0.324	1.06	16.98	0.394	0.345	10.119
UniTok	25.34	0.742	0.132	0.362	24.95	0.750	0.131	3.918
OmniTokenizer	24.69	0.771	0.138	1.411	24.31	0.779	0.137	6.292
Vila-U	22.24	0.612	0.228	4.231	21.89	0.620	0.227	10.997
OpenVision 3	30.92	0.902	0.053	0.187	30.89	0.907	0.050	1.601

3.3 Training settings

Training stages and resolution. In accordance with the conclusions drawn in CLIPA (Li et al., 2023b), we employ a progressive training strategy for the tokenizer, transitioning from low-resolution to high-resolution inputs. We first pre-train the tokenizer at 128×128, and then finetune it with 224×224 or 256×256. The epoch distribution for the two training stages is maintained at around a 10:1 ratio. By focusing most of the computation on low-resolution stages, our tokenizer attains superior performance while significantly reducing the computational overhead typically associated with high-resolution training.

Training details. As depicted in Figure 1, we use pre-trained FLUX.1 VAE and freeze it during the whole training process. All other components (including ViT encoder, ViT decoder, text encoder, text decoder, and linear layer) are randomly initialized and remain unfrozen throughout the training. For the two training stages, the global batch sizes are 8K and 4K, with cosine-decayed base learning rates of 8×10^{-6} and 4×10^{-7}. We disable LPIPS loss during the pretraining stage to prevent loss conflicts due to varying resolutions. For complete parameter and configuration details, please refer to Supplementary Section A.2. The model is trained on the DataComp dataset recaptioned by LLaVA-Llama-3 (Li et al., 2024b), which ensures the high quality of the training data and highly efficient multimodal learning.

4 Experiments

4.1 Evaluation settings

To comprehensively evaluate the performance of our unified tokenizer, we evaluate the reconstruction, generation and understanding performance and report their results in Section 4.2. For the generation side, we follow RAE configs to train a generative model with DiT and a wide DDT head, and evaluate the generation fidelity of OpenVision 3 on ImageNet. For the understanding side, we train vision-language models with our tokenizer under LLaVA-1.5 and LLaVA-NeXT frameworks (Liu et al., 2024a), and evaluate the understanding performance across a range of downstream multimodal benchmarks.

4.2 Reconstruction performance

As shown in Table 1, OpenVision 3 significantly outperforms existing unified tokenizers across all metrics. Previous unified models (Zheng et al., 2025; Ma et al., 2025; Wang et al., 2024; Wu et al., 2024) often struggle to maintain high reconstruction quality due to the trade-off required to align with semantic objectives (*e.g.*, SigLIP alignment). For instance, on ImageNet, OpenVision 3 achieves a PSNR of 30.92 dB, surpassing UniTok (25.34 dB) and Vila-U (22.24 dB) by a wide

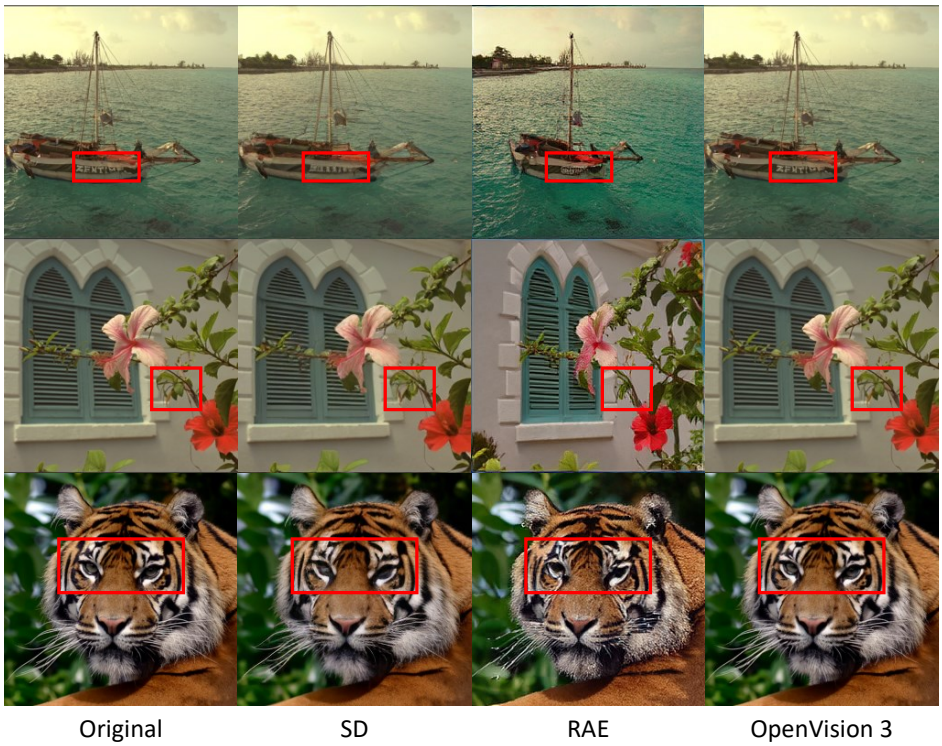


Figure 2: **Reconstruction visualization.** We present three cases of image reconstruction to compare OpenVision 3 with SD-VAE and RAE.

margin. Similarly, in terms of perceptual quality, our model achieves an LPIPS score of 0.053, whereas the closest unified competitor, UniTok, lags behind at 0.132. This demonstrates that our architecture—utilizing the VAE-ViT hybrid design—successfully mitigates the information loss typically associated with semantic compression. On the COCO dataset, our tokenizer keeps a substantial performance advantage. For instance, our rFID significantly outperforms UniTok (1.601 *vs*. 3.918) and other unified tokenizers by a large margin, further demonstrating our robust generalizability. Notably, even in comparison with specialized generation-oriented tokenizers (Rombach et al., 2022; Esser et al., 2024; Agarwal et al., 2025; Labs et al., 2025; Wan et al., 2025), our model maintains competitive or better results.

Reconstruction Visualization We present several qualitative results in Figure 2 to demonstrate our reconstruction performance. Compared to SD-VAE and RAE, OpenVision 3 exhibits superiority in preserving intricate image details and faithfully reconstructing text. In the first case, our tokenizer fully recovers the characters on the hull. In the second and third cases, our tokenizer exhibits superiority in capturing fine-grained details, such as the textures of flowers and foliage and the intricacies of the tiger's eyes, demonstrating a higher-fidelity reconstruction capability.

4.3 Generation Performance

We present our detailed generation results in Supplementary Section A.3. The generative performance of each tokenizer when paired with its respective compatible generator is shown in Table 11, where we report generation Fréchet inception distance (gFID), Inception Score (IS), Precision (Pre.), and Recall (Rec.). For low-level tokenizers, we evaluate SD-VAE with traditional diffusion-based generative models (DiT and SiT) (Peebles & Xie, 2023; Ma et al., 2024). For semantic tokenizers, we select CLIP, SigLIP2 and OpenVision with RAE generator for fair comparison with our tokenizer. According to Table 11, OpenVision 3 outperforms these tokenizers across all the metrics. For example, we achieve better gFID when compared to SD-VAE with improved generator SiT (**1.87 *vs*. 2.06**). Under the RAE frameworks, OpenVision 3 significantly outperforms CLIP and SigLIP2 (**1.87 *vs*. 2.27**). Our tokenizer also surpasses unified encoders like UniTok (Ma et al., 2025) by a large margin in generation (**1.87 *vs*. 2.51**).

Table 2: **Comparison of OpenVision 3 with OpenAI CLIP under LLaVA-1.5 framework.** We evaluate the understanding performance of OpenVision 3, CLIP, DFN-2B and LAION-2B.

Method	Vision Encoder	MME	Seed	SciQA	GQA	POPE	Chart	OCR	AI2D	MMStar	Avg.
OpenAI-CLIP	B/16	**1399/318**	62.2	**73.7**	58.6	82.9	**11.7**	143	**59.5**	**34.3**	50.2
OpenVision 3	VAE + B/2	1388/275	**63.1**	73.1	**58.9**	**83.7**	11.6	**172**	59.3	32.2	50.1
OpenAI-CLIP	L/14	**1468**/292	65.4	**73.9**	60.6	84.7	**12.6**	179	59.0	35.6	51.6
DFN-2B	L/14	1419/**325**	65.2	72.3	60.4	83.9	12.2	135	59.5	34.7	50.8
LAION-2B	L/14	1388/319	62.7	73.7	58.7	82.3	11.9	134	58.8	35.5	50.2
OpenVision 3	VAE + L/2	1405/313	**65.8**	72.4	**61.0**	**85.2**	11.9	**244**	**59.8**	**36.5**	52.4

Table 3: **Comparison of OpenVision 3 with OpenAI CLIP under LLaVA-NeXT framework.** We compare OpenVision 3 and OpenAI CLIP under LLaVA-NeXT setting.

Method	Vision Encoder	MME	Seed	SciQA	GQA	POPE	Chart	OCR	AI2D	MMStar	Avg.
OpenAI-CLIP	B/16	**1368**/316	61.2	**72.8**	58.1	84.7	32.0	**302**	**62.2**	32.2	54.1
OpenVision 3	VAE + B/2	1340/**337**	**63.3**	68.9	**59.2**	**84.9**	**35.0**	249	62.0	**33.1**	54.0
OpenAI-CLIP	L/14	1446/**319**	61.8	**75.3**	59.4	84.1	37.4	312	61.2	**34.9**	55.7
OpenVision 3	VAE + L/2	**1449**/280	**68.6**	73.6	**62.0**	**86.6**	**46.9**	**329**	**63.1**	33.4	57.9

Visualization. As shown in Figure 4, we visualize some class-conditional ImageNet-256 generation results with OpenVision 3 under RAE framework. The images generated by our tokenizer exhibit structurally coherent objects with rich stylistic details, which shows the strong capability of our tokenizer in generating high-quality samples with great fidelity and diversity.

4.4 Understanding performance

To evaluate the semantic representation capability of OpenVision 3, we integrate it into the LLaVA-1.5 and LLaVA-NeXT framework and train multimodal models. Due to the fixed downsample size of VAE, we keep the same image token numbers for fair comparison, where token number is 196 for B size and 256 for L size. In Table 2, we compare our tokenizer with CLIP, DFN, and LAION and present the results on multiple benchmarks. In Table 3, we further extend to LLaVA-NeXT framework to verify the robustness of OpenVision 3. We select MME (Fu et al., 2023), GQA (Hudson & Manning, 2019), MMStar (Chen et al., 2024) and SeedBench (Li et al., 2024a) for **comprehensive multimodal ability**; ScienceQA (Saikh et al., 2022) and AI2D (Kembhavi et al., 2016) for **science domain**; ChartQA (Masry et al., 2022) and OCRBench (Liu et al., 2024c) for **OCR domain**; POPE (Li et al., 2023c) for **hallucination evaluation**. Average scores are computed after normalizing each benchmark to a 0–100 scale.

Our tokenizer demonstrates strong performance and scalability. In Table 2, the Base-size OpenVision 3 achieves competitive performance with OpenAI CLIP (50.1 vs. 50.2), while showing advantage on specific domains (*e.g.*, 172 *vs*. 143 on OCR and 83.7 *vs*. 82.9 on POPE). When scaling to the Large size, our tokenizer outperforms CLIP by an average of **+0.8** points (**52.4 *vs*. 51.6**) and surpasses CLIP, DFN, and LAION on 6 out of 9 benchmarks. This scalability trend becomes even more pronounced under the LLaVA-NeXT framework. While the Base-size tokenizers still achieve comparable performance, OpenVision 3 obtains a **+2.2**-point lead at the Large size (**57.9 *vs*. 55.7**), and shows substantial advantages on specific benchmarks (*e.g.*, 46.9 *vs*. 37.4 on ChartQA and 68.6 *vs*. 61.8 on SeedBench). These results show that OpenVision 3 matches or outperforms understanding-oriented tokenizers in semantic capability across different frameworks. Our model also demonstrates strong scalability: compared with CLIP, our tokenizer achieves more substantial performance gains after scaling up the model size.

5 Discussion

5.1 Reciprocal synergy between understanding and reconstruction

For unified tokenizers, balancing the capabilities of understanding and generation remains a long-standing challenge. To investigate the interaction between understanding and reconstruction objectives within our tokenizer, we compare the full-loss tokenizer with variants trained using only

Table 4: **Understanding performance comparison between the full loss and only understanding loss variant.** We evaluate the two variants under the LLaVA-NeXT architecture.

Model	MME	Seed	SciQA	GQA	POPE	Chart	AI2D	MMStar	MMBench	Avg.
full loss	1367/**346**	60.6	**71.6**	57.1	83.6	28.4	**58.3**	29.9	**49.5**	**55.0**
only understanding loss	**1392**/302	**61.8**	70.4	**58.0**	**84.4**	**28.9**	56.1	**30.8**	48.5	54.7

Table 5: **Reconstruction and generation performance comparison between the full loss and only reconstruction loss variant.** Generation metrics are evaluated under the RAE framework. The RAE generator is trained for 8 and 10 epochs in stages 1 and 2, respectively.

Model	Reconstruction				Generation			
	PSNR↑	SSIM↑	LPIPS↓	rFID↓	gFID↓	IS↑	Precision↑	Recall↑
full loss	30.8	0.891	0.055	0.198	**45.82**	**23.15**	**0.42**	**0.53**
only reconstruction loss	**31.1**	**0.895**	**0.051**	**0.170**	74.09	17.03	0.31	0.45

the understanding loss or only the reconstruction loss. All three tokenizer variants follow the same training schedule of 400 epochs of pretraining followed by 100 epochs of finetuning.

Effect of reconstruction loss on understanding. In Table 4, we compare the full-loss tokenizer with the variant trained using only the understanding loss under the LLaVA-NeXT framework. Their understanding performance is broadly comparable, with gains and losses varying across benchmarks. The full-loss tokenizer achieves a slightly higher average score (55.0 vs. 54.7). These results suggest that adding reconstruction supervision does not compromise understanding but instead provides a modest overall benefit.

Effect of semantic loss on generation. In Table 5, we compare the full-loss tokenizer with the variant trained using only the reconstruction loss on reconstruction and downstream generation. The reconstruction-only variant performs better on all four reconstruction metrics, including PSNR (31.1 vs. 30.8) and rFID (0.170 vs. 0.198). However, when evaluated on downstream generation, the full-loss tokenizer achieves substantially better results, with a gFID of 45.82 vs. 74.09 and an IS of 23.15 vs. 17.03. Although adding semantic supervision does not directly improve reconstruction performance, it substantially enhances downstream image generation. Together, these results suggest a mutual benefit between the two objectives: reconstruction supervision modestly improves understanding, while understanding supervision substantially improves downstream generation.

5.2 VAE LATENTS HELP UNIFIED LEARNING

A core design of our tokenizer involves feeding VAE latents into a ViT for encoding. VAEs are originally designed for low-level visual encoding to capture pixel-level image features. However, we found that performing unified modeling with VAE latents substantially enhances image generation capabilities without compromising multimodal understanding performance, as shown in Table 6, Table 7 and Table 8. We argue that high-quality VAE latents enable a single ViT encoder to generate a single, unified visual representation that supports both generation and understanding. This avoids the need for separate encoding processes followed by concatenation or feature fusion. Furthermore, within this VAE latent space, we can achieve superior alignment between features for understanding and generation, leading to high performance in both capabilities.

To evaluate the necessity of VAE, we conducted a comparative analysis by training two tokenizers with raw image tokens and VAE latents, respectively. Specifically, since the downsample rate of FLUX-VAE is 8, we separately train VAE+ViT-B/2 and ViT-B/16 as the visual tokenizers to keep the same image token number, and all hyperparameters were kept strictly identical. For both sides, we pretrain at 128×128 resolution for 1000 epochs, and then finetune for 200 epochs at 256×256 for generation, and 224×224 for understanding.

In Table 6, we report the reconstruction performance with and without VAE. Compared to the standalone ViT, our tokenizer outperforms it by a large margin in rFID (0.216 *vs*. 0.980), demonstrating our high-quality reconstruction results. Meanwhile, there are marginal gaps in the other three metrics: PSNR, SSIM, and LPIPS. Furthermore, we train both tokenizers under RAE framework with

Table 6: **Effectiveness of VAE in reconstruction.**

Model	PSNR↑	SSIM↑	LPIPS↓	rFID↓
w/o VAE	32.82	0.935	0.060	0.980
OpenVision 3	30.33	0.885	0.061	0.216

Table 7: **Effectiveness of VAE in Generation.**

Model	gFID↓
w/o VAE	9.68
OpenVision 3	8.45

Table 8: **Understanding performance comparison of unified tokenizer with and without VAE.**

Method	Vision Encoder	MME	Seed	SciQA	GQA	POPE	Chart	OCR	AI2D	MMStar	Avg.
OpenVision 3	VAE + B/2	**1383/348**	**63.6**	**70.5**	**59.1**	**84.4**	31.3	**202**	60.7	**32.3**	53.2
w/o VAE	B/16	1346/296	62.0	67.8	57.7	83.8	**33.3**	193	**61.6**	**32.3**	52.2

Table 9: **Effectiveness of encoder size in reconstruction and generation.**

Model	Reconstruction				Generation			
	PSNR↑	SSIM↑	LPIPS↓	rFID↓	gFID↓	IS↑	Pre.↑	Rec.↑
OpenVision 3-B	30.92	0.902	0.053	0.187	8.45	144.7	0.78	0.36
OpenVision 3-L	30.96	0.902	0.052	0.186	8.89	148.3	0.78	0.36

100 epochs to study the effectiveness of VAE in generation. As shown in Table 7, our model achieves a gFID of 8.45 on ImageNet with the VAE, representing a **1.23 improvement** over the 9.68 gFID obtained without it. In Table 8, we evaluated the understanding capabilities of the tokenizer when the VAE is removed under the LLaVA-NeXT framework. The results show a modest advantage for OpenVision 3. Our tokenizer outperforms the counterpart by an average of **+1.0** point across 9 benchmarks (**53.2** ***vs.*** **52.2**). This suggests that although the VAE primarily extracts low-level visual features, it also contributes to the enhancement of multimodal understanding capabilities. These findings demonstrate that unifying tasks in the VAE latent space yields simultaneous gains in reconstruction fidelity, generative quality, and comprehensive understanding, validating the necessity of the VAE in our unified learning framework.

5.3 The selection of ViT encoder size

In Table 9, we explore the scaling behavior of the ViT encoder on the image reconstruction and generation. Generation results are evaluated under RAE pipeline with 100 training epochs. The results indicate that B-size and L-size tokenizers yield comparable results in reconstruction and generation. In contrast, Table 2 and Table 3 demonstrate that scaling up to L-size provides substantial gains in multimodal understanding. We argue that although increasing the size leads to a more representative tokenizer naturally, the outer VAE architecture imposes a theoretical ceiling on the generative performance. With sufficient training, the reconstruction performance of OpenVision 3 approaches that of FLUX-VAE (0.187 vs. 0.176 in rFID), and this bottleneck marginalizes the gains derived from larger encoders, presenting an open problem that requires further investigation in future research. We also provide more discussion on the selection of ViT decoder size in Supplementary Section A.1.

6 Conclusion

This work introduces OpenVision 3, a unified vision encoder for both understanding and generation. We innovatively couple a VAE with a ViT to form a unified architecture, and generate a single, unified representation for different downstream tasks. For the efficient training of our tokenizer, we propose a new training paradigm with both reconstruction- and semantics-driven signals for joint learning. Comprehensive evaluations reveal that our model yields superior results across generative and understanding tasks. OpenVision 3 outperforms other current unified tokenizers in reconstruction and generation, and shows superior ability compared with CLIP on semantic tasks. Furthermore, our architecture enables a mutual promotion relationship between image understanding and generation. The VAE component in our architecture design is proven to be critical through extensive experiments. To facilitate future research in the community, we will fully open-source our training code, data, and tokenizer checkpoints.

AI USE STATEMENT

The authors used AI as a writing assistant for grammar correction and sentence-level refinement, and GPT-4o is used as the evaluation judge, which is the standard pipeline for some public benchmark evaluation. All research contributions are entirely the authors' own work.

ETHICS STATEMENT

This work uses publicly available datasets and does not involve human-subject studies or the collection of new private data. As with other general-purpose vision encoders, potential biases in the training data and misuse in downstream image generation should be considered when applying the model.

REPRODUCIBILITY STATEMENT

The model architecture and learning objectives are described in the main text, with training and evaluation procedures in Sections 3.3 and 4.1 and full training configurations in Supplementary Section A.2. We will make anonymized source code, configuration files, all training and evaluation scripts, and model checkpoints publicly available. The recaptioned training data are publicly available.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

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? In *NIPS*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

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.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22487–22497, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

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

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

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024a.

Xianhang Li, Zeyu Wang, and Cihang Xie. Clipa-v2: Scaling clip training with 81.1% zero-shot imagenet accuracy within a $10,000 budget; an extra $4,000 unlocks 81.8% accuracy. *arXiv preprint arXiv:2306.15658*, 2023a.

Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. *Advances in Neural Information Processing Systems*, 36:49068–49087, 2023b.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024b.

Xianhang Li, Yanqing Liu, Haoqin Tu, and Cihang Xie. Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3977–3987, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023c.

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

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. Clips: An enhanced clip framework for learning with synthetic captions. *arXiv preprint arXiv:2411.16828*, 2024b.

Yanqing Liu, Xianhang Li, Letian Zhang, Zirui Wang, Zeyu Zheng, Yuyin Zhou, and Cihang Xie. Openvision 2: A family of generative pretrained visual encoders for multimodal learning. *arXiv preprint arXiv:2509.01644*, 2025a.

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*, 2024c.

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

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

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12490–12500, 2024.

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

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. In *IJDL*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

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

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

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

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

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

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

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

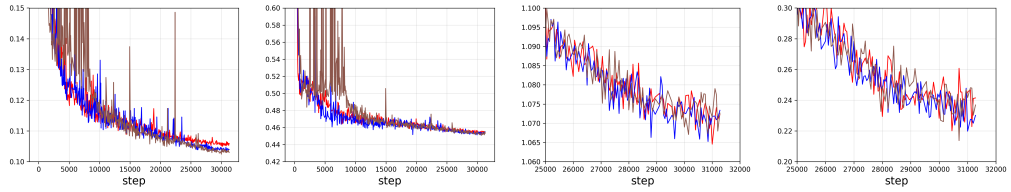
Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

TECHNICAL APPENDIX

A MORE DISCUSSION

A.1 THE SELECTION OF VIT DECODER SIZE

In the reconstruction branch, we use a ViT and a linear layer as the pixel decoder. In Figure 3, we investigate the impact of the size of the ViT decoder on reconstruction and understanding. Varying the ViT decoder variants for pixel-level decoding has minimal impact on the understanding branch. As illustrated in Figure 3c and Figure 3d, the semantic signals remain nearly invariant across different ViT decoder sizes. Regarding reconstruction learning, increasing the decoder size from M/1 to B/1 results in a significant reduction in pixel-level loss, as shown in Figure 3a. However, employing the larger L/1 variant leads to highly unstable training, as depicted by the brown curve. In terms of training time, we conduct experiments for 200 epochs for each size. The total training durations for the M/1, B/1, and L/1 decoder on a TPU v4-256 pod are 6.1, 6.0, and 6.9 hours, respectively. Taking into account both performance and computational efficiency, employing Base-sized ViT as the decoder strikes an optimal balance.



(a) pixel recon loss (b) latents recon loss (c) caption loss (d) contrastive loss

Figure 3: **Loss visualization with different ViT decoder sizes.** We use M/1, B/1 and L/1 variants as ViT decoder and draw the loss curves. As shown in Figures (a) and (b), employing a Large-sized decoder leads to significant training instability. While varying the decoder size has negligible impact on understanding performance, the Base-sized decoder effectively reduces reconstruction loss while maintaining training stability.

A.2 TRAINING CONFIGS

In Table 10, we present the detailed training parameters.

Table 10: **Parameter configs for two training stages.** The epoch number is defined as ImageNet-equivalent epochs (1 epoch $\approx$ 1.3M samples).

Parameter	Pretraining	Finetune
Resolution	128	224/256
Global batch size	8192	4096
Base learning rate	8×10^{-6}	4×10^{-7}
Epochs	4000	400
Warmup Epochs	40	20
LPIPS loss weight λ	0	0.5
VAE latents β	0.4	
Contrastive α	1.0	
Rec. loss ω_{rec}	0.5	
Und. loss ω_{und}	1.0	

A.3 Detailed results on Generation

In Table 11, we compare the generation performance of OpenVision 3 with other tokenizers with paired generators. In Figure 4, we show some qualitative results of OpenVision 3 on ImageNet-256 class-conditional generation. Our tokenizer is able to generate diverse and high-quality images.

Table 11: **Class-conditional image generation on ImageNet 256x256.** We report gFID, Inception Score (IS), Precision (Pre.), and Recall (Rec.). Our tokenizer achieves the best performance across all metrics.

Tokenizer	Generator	gFID↓	IS↑	Pre.↑	Rec.↑
SD-VAE	DiT	2.27	278.2	0.83	0.57
SD-VAE	SiT	2.06	270.3	0.82	**0.59**
UniTok	LlamaGen	2.51	216.7	0.82	0.57
CLIP	RAE	2.54	256.4	0.80	0.54
SigLIP2	RAE	2.27	267.2	0.80	0.54
OpenVision	RAE	2.44	262.2	0.80	0.53
OpenVision 3	RAE	**1.87**	**290.0**	**0.84**	**0.59**



Figure 4: **Qualitative results of ImageNet class-conditional generation.**