Towards Scalable Pre-training of Visual Tokenizers for Generation

Jingfeng Yao[1⋆], Yuda Song[2], Yucong Zhou[2], and Xinggang Wang[1⋆⋆]

[1] Huazhong University of Science and Technology, Wuhan, China
[2] MiniMax, Shanghai, China

Abstract. The quality of the latent space in visual tokenizers (e.g., VAEs) is crucial for modern generative models. However, the standard reconstruction-based training paradigm produces a latent space that is biased towards low-level information, leading to a foundational flaw: better pixel-level reconstruction accuracy does not lead to higher-quality generation. This implies that pouring extensive compute into visual tokenizer pre-training translates poorly to improved performance in generation. We identify this as the "pre-training scaling problem" and suggest a necessary shift: to be effective for generation, a latent space must concisely represent high-level semantics. We present **VTP**, a unified visual tokenizer pre-training framework, pioneering the joint optimization of image-text contrastive, self-supervised, and reconstruction losses. Our study reveals that **perception-oriented tokenizer pre-training unlocks a new scaling law for generation**, where generative performance scales effectively with compute, parameters, and data allocated to the pre-training of the visual tokenizer. Our large-scale pre-training experiments demonstrate the following results: (1) Without modifying DiT training specs and FLOPs, solely scaling VTP pre-training consistently achieves gains in both ImageNet class-conditional and LAION text-to-image generation, while conventional autoencoders stagnate very early at 1/10 of the FLOPs. (2) VTP achieves 0.36 rFID while simultaneously delivering 78.2% zero-shot accuracy and 85.7% linear probing accuracy, surpassing prior unified tokenizers such as VILA-U and UniTok. (3) Furthermore, the VTP-based diffusion model exhibits exceptionally fast convergence—reaching 2.03 gFID in only 80 epochs without guidance tricks, outperforming previous methods like VA-VAE and RAE—and ultimately scales to achieve a remarkable *1.11 gFID* on ImageNet 256×256 generation. Our code and models are publicly available at VTP.

Keywords: Visual Tokenizer · Diffusion Model · Pre-training

1 Introduction

Latent Diffusion Models (LDMs) [28] employ a visual tokenizer, such as a VAE [16], to compress visual signals into a latent space. Typically, visual tokenizers are pre-trained in a separate stage using a reconstruction objective.

⋆ This work was completed while Jingfeng Yao was an intern at MiniMax.
⋆⋆ Corresponding author: xgwang@hust.edu.cn

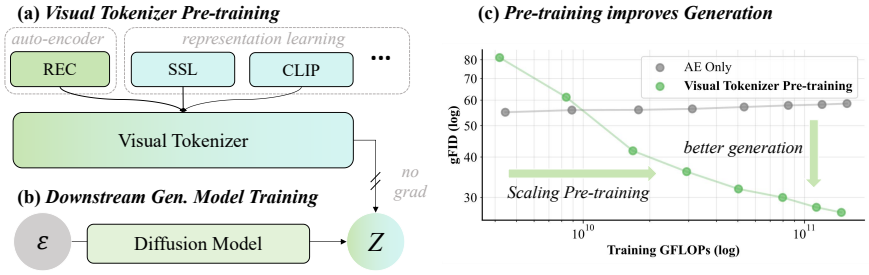


Fig. 1: Visual Tokenizer Pre-training. We revisit the visual tokenizer pre-training in LDM [28] from a representation learning perspective. Critically, while keeping the diffusion model (e.g., DiT [24]) training configuration and FLOPs fixed, our method improves generation solely by scaling the tokenizer's pre-training to learn a better-structured latent space.

However, a clear paradox has emerged: better reconstruction does not guarantee better generation. Instead, a noticeable trade-off between the two objectives is widely observed [9,37,39]. It implies that scaling the computational investment in pre-training, while potentially further improving reconstruction performance, carries the risk of compromising generation performance (see Fig. 1 (c)), which is consistent with prior work [11]. We note that this limitation of reconstruction-only training may arise because the objective biases the latent space toward low-level information and, as training scales up, increasingly drives it away from the structured latent space we ultimately desire, thereby motivating the search for tokenizer pre-training schemes that genuinely scale—a challenge we term the *"pre-training scaling problem"*.

Unlike conventional approaches that emphasize low-level information, we propose that an effective latent space for generation should efficiently encode the core visual semantics. Early explorations have already demonstrated the value of this principle through two primary pathways. Some works explicitly enrich the latent space with specific semantic objectives—for instance, by concatenating optical flow for motion or leveraging powerful pre-trained features to the latent space, like VideoJAM [3] and ReDi [18]. Others implicitly structure the space using semantic constraints, an approach seen in methods like VA-VAE [39] and REPA-E [19], which regularize the VAE's feature space with representational priors. While promising, these efforts remain preliminary and do not explore broader scaling properties.

To address this challenge, we present **VTP**, a novel pre-training framework for visual tokenizers. The core contribution of our work is a redesigned, scalable paradigm for visual tokenizer pre-training that benefits generation. This is achieved by jointly optimizing the model across a spectrum of visual representation tasks, including cross-modal alignment, global semantic understanding, local spatial perception, and low-level pixel reconstruction. Technically, our framework is built upon a vision transformer (ViT) [8] based Auto-Encoder. Building on

the flexibility of the ViT architecture for representation learning, we integrate a suite of diverse learning objectives. First, cross-modal image-text contrastive learning is employed to instill a global semantic understanding [15, 26]. This is complemented by integrating established self-supervised learning techniques, notably self-distillation and mask image modeling [2, 12, 23, 45], to enhance the model's spatial-semantic perception. Throughout this multi-task learning process, the pixel-level reconstruction objective is consistently applied to preserve fine-grained visual details for generation. We posit that this holistic training paradigm encourages the latent space to form a unified and rich representation of visual information, which is instrumental in boosting the fidelity and semantic coherence of the generated outputs. (Sec. 3)

Through extensive experiments, we reveal that *perception-oriented tokenizer pre-training unlocks a new scaling frontier for generation*: (1) Understanding is a key driver of generation: The introduction of semantic understanding and perception tasks enhances the generative capability of models initially pre-trained solely on reconstruction. We observe a strong positive correlation between the semantic quality of the latent space and its generative performance. While these tasks differ in paradigm, they consistently inject more meaningful representations, leading to significant gains in downstream generation. (see Fig. 2) (2) Superior Scalability for Generation: VTP is the first visual tokenizer to demonstrate scaling properties. Its generative performance improves steadily as we scale up training compute (FLOPs), model parameters, and dataset size of the visual tokenizer. This stands in stark contrast to traditional tokenizers pre-trained only on reconstruction, whose performance rapidly saturates and shows negligible gains with increased scale.

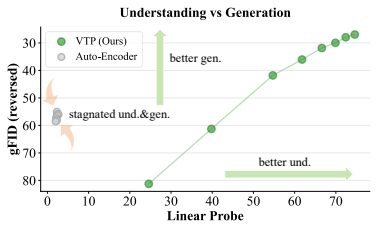


Fig. 2: Understanding is a key driver of generation. We observe a strong positive correlation between the comprehension and generative capabilities of the latent space during visual tokenizer pre-training.

We conduct extensive experiments on ImageNet [7] class-conditional generation and LAION [29] text-to-image generation. (1) We demonstrate new scalability for visual tokenizer pre-training: increasing pre-training compute accelerates the convergence of downstream VTP-based diffusion models. Specifically, scaling compute by 10× yields a 65.8% FID improvement for DiT on ImageNet, overcoming the early saturation of conventional autoencoders. (2) A systematic analysis reveals that diverse perception losses consistently improve general generation quality (Sec. 4), and (3) specific losses offer targeted benefits for downstream tasks; *e.g.*, CLIP loss significantly enhances text-to-image generation (Sec. 5). (4) Our final model gets a strong performance, achieving 0.36 rFID on ImageNet alongside 78.2% zero-shot and 85.7% linear probing accuracy, outperforming unified tokenizers like VILA-U [35] and UniTok [21]. Coupled with a DiT [24, 39], it enables exceptionally fast convergence (2.03 FID in just 80 epochs, unguided),

surpassing prior approaches such as VA-VAE [39] and RAE [44], and ultimately reaches a remarkable 1.11 gFID on generation tasks.

To sum up, our contribution could be summarized as follows:

- We propose Visual Tokenizer Pre-training (VTP), a framework integrating contrastive, self-supervised, and reconstruction objectives to build a perception-oriented tokenizer pre-training for generative models.
- We demonstrate a new scaling property for visual tokenizers: overcoming the early saturation of reconstruction-only training, downstream generation quality consistently improves as VTP pre-training scales in compute, parameters, and data.
- VTP pushes the boundaries of unified understanding and generation of visual encoder, achieving an impressive 1.11 gFID on ImageNet generation alongside 0.36 rFID, 78.2% zero-shot, and 85.7% linear probing accuracy.

2 Related Work

2.1 Pre-training and Representation Learning

Pretraining is typically a scalable paradigm for boosting downstream task performance by first optimizing models on large-scale data with specific objectives. The early paradigm relied on supervised pretraining—such as ImageNet classification [8, 13]—and transferred weights to downstream tasks like detection [27] and segmentation [36]. A recent paradigm shift has focused on weakly-supervised and unsupervised methods to enable pretraining at larger scales. For instance, CLIP [26] uses image-text contrastive learning with weak supervision by minimizing the distance between image and text features. SigLIP [41] further optimizes this process via a sigmoid loss for large-scale training. Another branch, self-supervised learning (SSL), learns directly from unlabeled data. Methods like MAE [12] and BEiT [1] adopt masked image modeling (MIM), training models to reconstruct masked patches. DINO [2] utilizes self-distillation to enforce multi-view classification consistency. iBOT [45] and DINOv2 [23] combine MIM with self-distillation to learn more generalized representations.

Despite these advances, within the explicitly decoupled, two-stage framework of LDMs [28]—comprising a visual tokenizer followed by a generative model—how to pretrain the first-stage tokenizer to enhance second-stage generative performance has not been systematically explored.

2.2 Diffusion with Pre-trained Representations

Recent works leverage pretrained visual representations to improve diffusion model training. For example, REPA [40] aligns DiT features with pretrained representations during training. DDT [34] decouples the alignment process into a condition encoder and a velocity decoder. iREPA [31] studies global information and spatial structure in representation alignment.

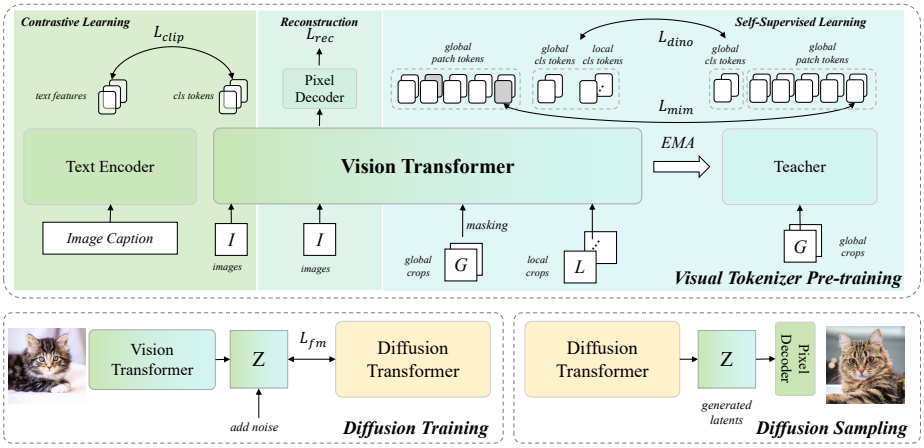


Fig. 3: Overview of Visual Tokenizer Pre-training (VTP). By integrating representation learning (image-text contrastive [26] and self-supervised learning [23]) with reconstruction within a Vision Transformer Auto-Encoder, we find that VTP exhibits a **well-behaved scaling property** for generative performance.

Another group of work has explored the use of visual representations to structure the latent space, which falls into two categories. The first employs a distillation objective: VA-VAE [39] aligns its latent space with features of visual foundation models to alleviate the trade-off between reconstruction and generation. ImageFolder [20] decouples semantic and pixel-level feature spaces to improve autoregressive generation. MAETok [5] enhances latent representations by incorporating DINOv2 features into its MIM pre-training objective. REPA-E [19]concurrently optimizes the feature space of the VAE during DiT training by leveraging supervision from a pre-trained foundation model. I-DeTok [38] improves the latent space's suitability for both autoregressive and diffusion models generation via a joint pre-training strategy that employs MIM and noise injection. The second strand directly utilizes pre-trained representations for generation. AlignTok [4] adapts pretrained vision encoders into tokenizers for latent diffusion models. VFMTok [43] reorganizes frozen vision features for autoregressive image generation. Further, BLIP3-o [6] directly regresses SigLIP features and employs an SD-XL-based [25] decoder to boost efficiency. Recently, RAE [44] leverages DINOv2 features and trains a separate pixel decoder for reconstruction. Complementary to representation-based methods, EQ-VAE [17] and SER [32] regularize autoencoder latents with transformation-equivariance constraints.

However, the pre-trained representation-based methods are inherently limited by existing foundational models, often leading to a low performance ceiling or substantial reconstruction loss. Concurrently, while previous studies achieved performance improvements under specific configurations, the scalability of the proposed methods generally remains unverified.

3 Visual Tokenizer Pre-training

Our work introduces a scalable visual tokenizer pre-training paradigm that benefits generation. To this end, we integrate representation learning objectives with the conventional reconstruction loss to learn visual representations that are semantically rich, accurate in reconstruction, and generation-friendly (see Fig. 3).

3.1 Architecture

Leveraging its flexibility in learning visual representations, our visual tokenizer uses a fully Vision Transformer (ViT) architecture. In line with standard autoencoder designs, we introduce a bottleneck that maps visual information into a d-dimensional latent space. Encoder features are leveraged by the text encoder, EMA teacher, and pixel decoder to facilitate their distinct training objectives.

3.2 Visual Reconstruction

Given an image $I \in \mathbb{R}^{3\times H\times W}$, we compress it into a latent space $\mathbb{R}^{d\times H/16\times W/16}$ using a visual tokenizer and subsequently reconstruct into I' via a pixel decoder, which lifts latents back to the feature space, refines them with N ViT blocks, and reconstructs images in pixel space through a final pixel-shuffle layer.

The reconstruction task is challenged by the poor compatibility of GAN loss [10] with the ViT architecture, which causes large gradient norms and low training stability. To address this, we employ a two-stage training strategy. In the first stage (i.e., the pre-training stage), all parameters are jointly optimized by minimizing a composite loss function comprising the $\mathcal{L}_1$ loss and a perceptual loss $\mathcal{L}_{\text{perceptual}}$ [42] between I and I'. During the second stage, the visual tokenizer remains frozen while the pixel decoder is fine-tuned with a GAN objective to improve fidelity.

The overall reconstruction loss $\mathcal{L}_{\text{rec}}$ during pre-training is defined as:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_1 + \mathcal{L}_{\text{perceptual}} \tag{1}$$

3.3 Self-Supervised Learning

Following DINOv2 [23], our self-supervised learning framework comprises two components: masked image modeling (MIM) [12, 45] and self-distillation [2].

For a given image I, we apply data augmentation to obtain global and local views I_{global} and I_{local}. In MIM, adhering to [45], I_{global} is patch-embedded and fed directly to an EMA teacher, while its masked version is processed by the visual tokenizer, optimizing the complementary masking loss $\mathcal{L}_{\text{mim}}$. For self-distillation, similar to [2], I_{global} and I_{local} are passed to the visual tokenizer, and I_{global} to the EMA teacher, with the cross-entropy loss $\mathcal{L}_{\text{dino}}$ applied to their pseudo-label predictions.

Therefore, the overall self-supervised learning loss is defined as:

$$\mathcal{L}_{ssl} = \mathcal{L}_{mim} + \mathcal{L}_{dino} \tag{2}$$

3.4 Contrastive Learning

Given a batch of image-text pairs, we encode the image I and text T using a visual tokenizer and a text encoder, respectively, to obtain their visual and textual features. Following CLIP, we then maximize the similarity of the corresponding (positive) image-text pairs while minimizing the similarity of the remained non-corresponding (negative) pairs. This objective is formulated as the contrastive loss $\mathcal{L}_{\text{clip}}$.

3.5 Overall Objective

Building upon the preceding components, we integrate them into a unified pre-training framework. The overall training objective for our visual tokenizer pre-training is formulated as a weighted combination of the aforementioned losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ssl}}\mathcal{L}_{\text{ssl}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} \tag{3}$$

where $\lambda_{\text{rec}} > 0$, $\lambda_{\text{ssl}} \geq 0$, and $\lambda_{\text{clip}} \geq 0$ are balancing coefficients that control the contribution of each objective. This multi-task learning scheme enables the model to concurrently develop high-fidelity reconstruction capability, semantically rich representation learning, and cross-modal alignment, thereby establishing a robust and scalable visual tokenizer for diverse generation tasks.

3.6 Batch Sampling

We observe a significant disparity in optimal batch sizes across different training paradigms. Contrastive learning frameworks like CLIP demand extremely large batches (*e.g.*, 16k or 32k), while self-supervised and reconstruction objectives are typically effective with orders of much smaller batches (*e.g.*, 4k).

Given an input batch of B image-caption pairs, all samples are used for CLIP training, e.g. $B_{\text{clip}} = B$. B_{ssl} and B_{rec} are random sampled from B to accommodate the divergent batch size requirements of self-supervised learning and reconstruction.

4 Experiments

4.1 Implementation Details

Pre-training Our model architecture builds upon the Vision Transformer (ViT) implemented in [30]. We incorporate QKNorm [14] to enhance training stability. We employ a 12-layer transformer with a hidden dimension of 768 as the text encoder, and a 4-layer ViT-Large layer as the pixel decoder for fast experimentation. In designing the latent bottleneck, we primarily adopt a dimension of 64, following [21], to balance semantic comprehension with reconstruction quality. An ablation study on this configuration is conducted by varying the dimension

Arch.	FLOPs	#Params	rPSNR	gFID
CNN [28]	389.4G	70.3M	30.63	59.53
ViT-B [8]	87.7G	171.2M	30.72	58.40
ViT-L [8]	311.1G	607.2M	**31.28**	**53.51**

Table 1: AutoEncoder Performance with Different Architectures. ViT serves as a comparable alternative to CNNs, enabling simpler pre-training scaling experiments.

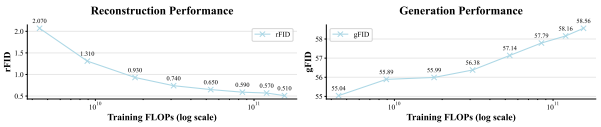


Fig. 4: Reconstruction Only Training Target CANNOT Lead to Effective Scaling for Downstream Diffusion Models. As training progresses, the tokenizer's reconstruction performance improves, while its generative performance degrades concurrently. It reveals the *inadequacy of pure reconstruction tasks for scalable tokenizer pre-training.*

to 256. We use an internally filtered version of DataComp-1B [7] with 277M samples for tokenizer pretraining, and ImageNet [7] for downstream DiT training. We set $B_{\text{clip}} = 16k$, $B_{\text{ssl}} = 4k$ and $B_{\text{rec}} = 2k$. For weighting, we set $\lambda_{\text{rec}} = 0.1$, while λ_{clip} and λ_{ssl} are set to either 0 or 1. We find that a smaller reconstruction weight contributes to improved generative performance. For self-supervised and contrastive pretraining implementation, we closely follow the established practices of DINOv2 [23] and OpenCLIP [15].

Downstream DiT training & evaluation We train the Diffusion Transformer (DiT) [24] under a fixed configuration to evaluate the generative capability of our visual tokenizer. Specifically, we follow LightningDiT [39] as a strong baseline. We report FID-10k scores obtained with a LightningDiT-B [39] model trained on ImageNet [7] for 80 epochs under a consistent protocol. For the reconstruction evaluation, the performance of all tokenizers is assessed on the standard ImageNet validation set at a resolution of 256. For most cases, we report rFID as the reconstruction metric. For understanding evaluation, we evaluate representation performance on ImageNet using linear probing. We do not employ the feature enhancements common in the DINO [23, 30] series, which can substantially increase linear probing scores by leveraging multi-layer features. Instead, we probe only the reduced-dimensionality features from the bottleneck, thereby directly evaluating the inherent properties of the latent features, and report ImageNet Top-1 acc. as the understanding metric.

4.2 Auto-Encoder with Vision Transformers

We begin by demonstrating that a Vision Transformer (ViT) can serve as an effective substitute for CNNs in standard reconstruction tasks. We construct a ViT visual tokenizer with a symmetric encoder and decoder. It has the specification of f16d64, where 'f' denotes downsample ratio (or patch size for ViT) and 'd' denotes bottleneck dimension. A two-stage training pipeline discussed in Sec. 3.2 is adopted to enhance training stability. Then, we implement the convolutional LDM architecture [28] under the same specifications. We use ImageNet at 256 resolution for training and testing.

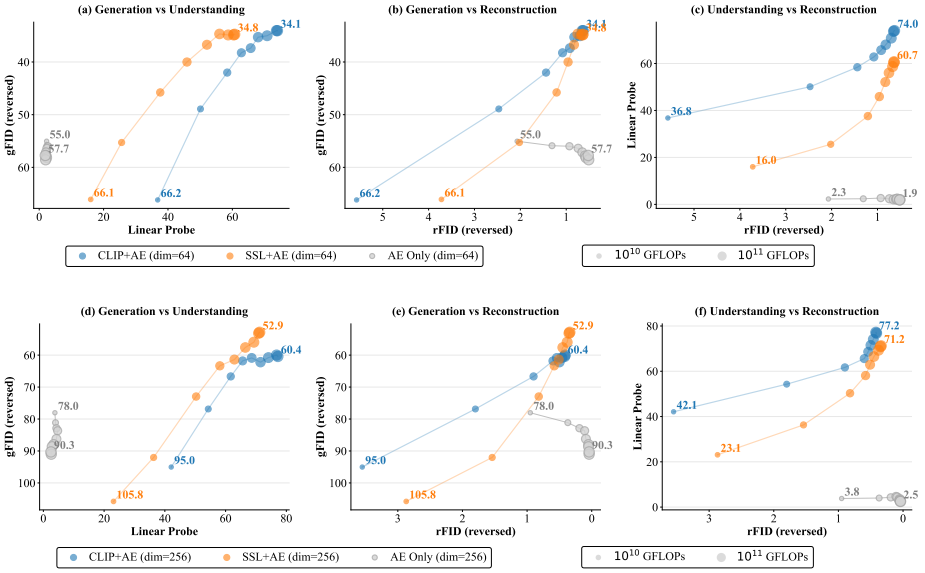


Fig. 5: Scalability of CLIP+AE & SSL+AE Visual Tokenizer Pre-training. Scaling properties under different strategies and bottleneck dimensions. Our method shows correlated growth in generation and comprehension with compute, while VAE-based tokenizer performance rapidly saturates.

As illustrated in Tab. 1, we evaluate their reconstruction and generation performance, observing that ViT-L achieves a reconstruction PSNR of 31.28 and a gFID of 53.51, on par with LDM. While it utilizes more parameters, it requires lower computational cost. These findings are consistent with previous observations [11, 33], suggesting that the simple design of this ViT tokenizer architecture is effective.

4.3 Scaling up Visual Tokenizer Pre-training

Our work focuses on how to scale up the tokenizer pre-training to improve the model's capabilities for downstream generative training.

We conduct three distinct scaling-up experiments. For these experiments, we employ a ViT-L backbone as the encoder and a lightweight decoder composed of 4 ViT-L layers to facilitate rapid training and inference. All models are trained on the 277M DataComp-filtered dataset introduced above.

Scaling with reconstruction only CANNOT help generation. Initially, we scale up the training computation for a standard reconstruction tokenizer. As illustrated in Fig. 4, we observe a scaling paradox: the model's reconstruction performance improves substantially with increased training compute, with rFID improving

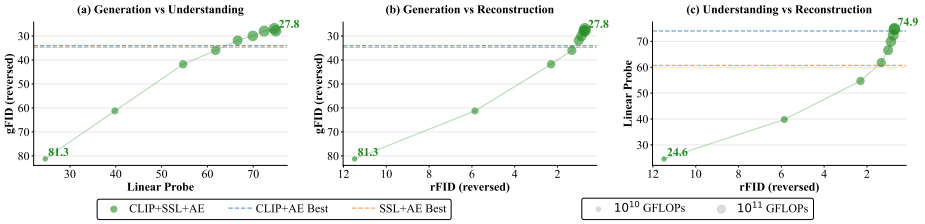


Fig. 6: Scalability of CLIP+SSL+AE Visual Tokenizer Pre-training. Under the same computational budget, the f16d64 tokenizer trained with joint CLIP and SSL representation learning achieves the best performance in both generation and comprehension.

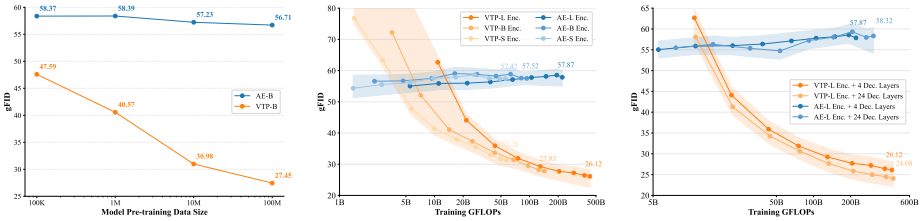
from 2.0 to 0.5. However, its generative performance in fact slightly degrades, as indicated by the gFID rising from 55.04 to 58.56. We posit that this phenomenon occurs because the reconstruction objective effectively guides the model to capture low-level details but provides insufficient incentive for learning high-level semantic representations, which are crucial for generation. Reconstruction task itself does not exhibit scalability in pretraining for downstream generation.

Scaling with different understanding tasks helps generation in a similar way. Then, we scale up visual tokenizer pre-training with the assistance of representation tasks. As described in Sec. 3, we integrate the reconstruction task with either image-text contrastive learning (CLIP) [26] or self-supervised learning (SSL, specifically DINOv2) [23] in a joint training framework. These two hybrid approaches are denoted as CLIP+AE and SSL+AE, respectively. To further substantiate the robustness of our conclusions, we include an additional experimental configuration with a latent dimension of $d = 256$. For a fair comparison, all Autoencoders (AEs) across different latent dimensions were trained under an identical computational budget. We concurrently monitor four key metrics: understanding performance, reconstruction fidelity, generation quality, and training FLOPs for a comprehensive evaluation.

Our experimental results, summarized in Fig. 5, lead to the following key observations:

(1) *Feasibility of Hybrid Objectives*: Hybrid training combining representation learning with reconstruction is viable. As evidenced by the Fig. 5 (c)&(f), traditional autoencoders (AEs) trained solely on reconstruction maintain low understanding performance. In contrast, when augmented with representation learning objectives—either CLIP or SSL—both understanding and reconstruction metrics exhibit stable, simultaneous improvement.

(2) *Negative Impact of Pure Reconstruction*: Solely relying on reconstruction proves counterproductive for downstream generation tasks. The Fig. 5 (b)&(e) illustrates a negative yield in reconstruction-only AEs: as the computational budget increases, reconstruction performance improves but the generation performance degrades.



(a) Data Scaling. **(b) Encoder Scaling.** **(c) Decoder Scaling.**

Fig. 7: Scalability of data and parameters. We observe a new scaling property: the DiT generation performance within fixed training FLOPs increases while its tokenizer has larger model sizes and training data.

(3) *Understanding as the Key Driver*: The integration of semantic understanding tasks counteracts this negative effect and emerges as the dominant factor for improving generation. The reversal of this trend is visible in the Fig. 5 (a)&(d) and (b)&(e). Specifically, our visual tokenizer pre-training, which jointly optimizes for reconstruction and representation learning, enables continuous, concurrent improvement in reconstruction, understanding, and generation as pre-training scales. Conversely, focusing exclusively on reconstruction optimization yields superior reconstruction, but leads to the stagnation of both understanding and generation performance. These observations imply that understanding is the key driving force necessary for effective generation.

(4) *Generality of Representation Learning*: Diverse representation learning paradigms, including CLIP and SSL, consistently enhance generation performance. Despite significant differences in their training frameworks, they share a critical mechanism: enriching the semantic understanding within the latent space. Although their scaling behaviors differ slightly, both methods substantially improve the efficacy of visual tokenizer pre-training for downstream generative tasks. This observation also suggests that new and emerging representation learning techniques can be seamlessly integrated to establish even better performance bounds.

Scaling with multiple understanding tasks gets better performance. Subsequently, we introduce, to the best of our knowledge, the first integration of contrastive, self-supervised, and reconstruction objectives (CLIP+SSL+AE) for visual tokenizer pre-training. Our experiments demonstrate that this training paradigm is feasible and stable. This multi-objective framework enables the tokenizer to capture multi-scale features, enhancing both semantic alignment and spatial fidelity. As shown in Fig. 6, our method under the f16d64 setting achieves a higher generative upper bound (gFID=27.8) alongside better understanding performance (74.9% linear probing accuracy) under a fixed computational budget. All subsequent data and parameter scaling experiments are based on this pre-training configuration.

4.4 Scaling Properties with Parameters

VTP demonstrates a new interesting parameter scalability, as its downstream DiT generative performance improves consistently with increased previous tokenizer model size.

We first investigate encoder scaling by training three ViTs of varying sizes using CLIP+SSL+AE and a baseline AE. As shown in Fig. 7 (b), the generative performance of the AE remains stagnant at about 57, regardless of the model capacity (from 20M to 300M parameters). In contrast, VTP exhibits a clear scaling trend: its gFID improves steadily from 31.28 to 26.12 as the model size grows, forming a well-defined parameter scaling curve. We then proceed to scale up the pixel decoder. Our findings indicate that this architectural expansion also leads to a correlated improvement in generative performance, with the gFID score decreasing from 26.12 to 24.08.

4.5 Scaling Properties with Data

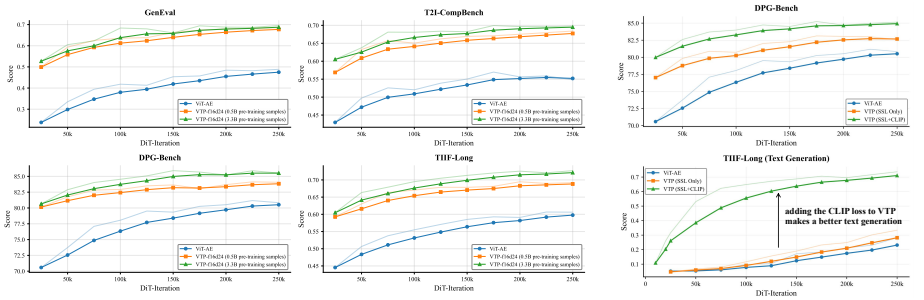
The scale of training data is also crucial for the generalization ability of the tokenizer. To validate this, we constructed four subsets of varying scales—100K, 1M, 10M, and 100M—by randomly sampling from the Datacomp-1B dataset. We trained both VTP-ViT-Large and AE-ViT-Large architectures on these subsets for 1.1 billion samples each and evaluated their generation performance. The results are summarized in Fig. 7 (a).

We draw the following observations from the results: First, VTP consistently outperforms the conventional autoencoder across all data scales. More importantly, the generative performance of the autoencoder shows negligible improvement with increased data, with its FID score merely decreasing from 58.37 to 56.71. In stark contrast, the performance of VTP improves significantly as the volume of training data grows, with its FID score substantially improving from 47.59 to 27.45. Notably, the downstream DiT training FLOPs remained strictly identical. This compellingly demonstrates that the introduced representation learning effectively enhances the data scalability of the visual tokenizer.

5 Scaling to Text-to-Image Generation

To validate the generalizability of VTP's scaling properties beyond class-conditional generation on ImageNet, we extend our evaluation to text-to-image (T2I) generation tasks on the LAION dataset [29].

We train VTP tokenizers and evaluate T2I generation performance across different pre-training compute budgets. For the downstream generative model, we adopt a DiT-XL architecture for generation. During training, images are resized with a short edge randomly sampled between 256 and 512. The results are shown in Fig. 8a. Additional details are provided in the supplementary material.



(a) VTP scaling properties generalize to T2I generation. VTP pre-trained with representation learning objectives consistently improves with increased compute, achieving better generation performance than reconstruction-only tokenizers.

(b) Ablation on Semantic losses. Adding CLIP loss to VTP significantly improves the text rendering ability in T2I.

Fig. 8: Text-to-Image Experiments on LAION.

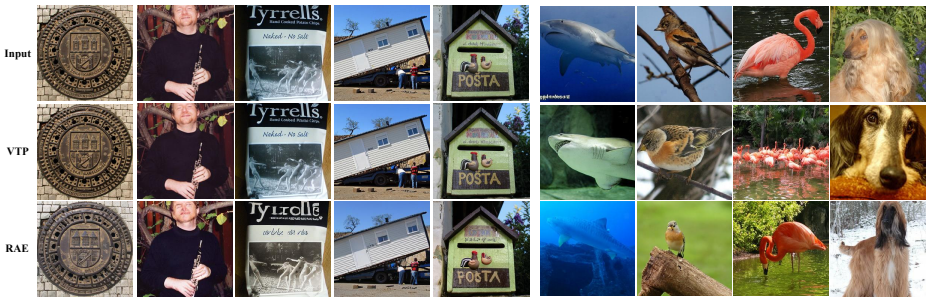
VTP scaling properties generalize to T2I. We observe two key findings: (1) Tokenizer with representation learning objectives achieves significantly faster convergence than the reconstruction-only AE baseline, demonstrating that semantic-aware tokenizer pre-training substantially improves training efficiency for downstream T2I generation. (2) As the tokenizer pre-training compute increases, the downstream generative model continues to improve, confirming that VTP's scaling properties are not limited to class-conditional ImageNet generation but generalize to the more challenging T2I setting.

CLIP loss improves text rendering. We further ablate the effect of different semantic losses on T2I generation quality (Fig. 8b). Similar to the results in Fig. 6 on ImageNet, as the variety of perception losses incorporated into VTP pre-training increases, we observe a steady improvement in downstream T2I generation performance. Furthermore, we observe that the tokenizer trained with the CLIP loss exhibits a significant advantage in text rendering for text-to-image generation, substantially outperforming both AE and SSL+AE and tokenizers.

6 Further Scaling and Discussion

We further scale VTP pre-training with larger training compute across three symmetric ViT architectures (VTP-S, VTP-B, VTP-L) and summarize results in Tab. 2 and Tab. 4.

Scalability Compared to Fixed Representation AutoEncoders. Compared to concurrent work RAE [44]. A distinctive advantage of VTP is its scalability. As shown in Table 2, under identical DiT training configurations, VTP's downstream generation improves consistently as the tokenizer scales from S to L, whereas RAE's performance degrades at larger scales. Moreover, since VTP always involves reconstruction during training, it preserves fine-grained details significantly better than RAE (see Fig. 9a).



(a) Reconstruction Comparison with RAE. Visual tokenizer pre-training with reconstruction target enables better reconstruction capacity than direct representation encoder transfer.

(b) Convergence on ImageNet. With 80 epochs of training, VTP achieves 2.61 gFID, significantly outperforming VA-VAE (4.29) and RAE (4.28).

Fig. 9: Visualizations of Reconstruction and Generation.

Tok.	RAE	VTP
S	3.50	5.46
B	4.28	3.88
L	6.09	**2.81**

Table 2: Scalability comparison. LightningDiT gFID↓ at 80 epochs.

	AE	×1/4	×1/2	×1	×2
CLIP	34.57	31.89	–	29.43	**27.83**
Rec	34.57	–	30.34	29.43	28.96

Table 3: Batch-sampling ablation. Generation FID↓ under different CLIP/rec batch allocations.

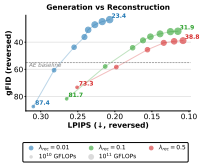


Fig. 10: Loss-balance ablation. Varying λ_{rec}.

Unified Performance Frontier. Our final model achieves 0.36 rFID, 78.2% zero-shot accuracy, and 85.7% linear probing accuracy on ImageNet, surpassing prior unified tokenizers VILA-U [35] and UniTok [21].

Superior Generation Without Architecture Modification. Built upon the standard LightningDiT [39], VTP-L achieves superior 1.11 gFID with guidance, surpassing all prior methods (Tab. 4). Moreover, VTP attains 2.60 and 2.03 gFID without guidance in only 80 epochs, demonstrating remarkably fast convergence. Additional details are provided in the supplementary material.

Discussion. We provide discussions on the loss weight and batch sampling strategy of VTP. For the loss weight, we adopt the CLIP+AE setting, fix λ_{clip}, and vary λ_{rec} to balance understanding and reconstruction. As shown in Fig. 10, the compute-scaling property holds across all λ_{rec} settings: downstream gFID consistently improves as tokenizer pre-training compute increases. Meanwhile, smaller λ_{rec} values yield better generation quality at the cost of reconstruction fidelity. We adopt λ_{rec}=0.1 for final training to achieve a good trade-off among perception, generation, and reconstruction. For batch sampling, we adopt the CLIP+AE setting with default batch sizes of 8k for CLIP and 1k for reconstruction, and vary each independently. As shown in Tab. 3, all batch sampling

Method	Tokenizer			Gen Model Params	Epochs	Generation w/o guidance				Generation w/ guidance			
	rFID↓	Zero-Shot↑	Linear Probe↑			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Perception or Unified Tokenizer Baselines													
SigLIP [41]	-	80.5	-	-	-	-	-	-	-	-	-	-	-
MAE [12]	-	-	85.9	-	-	-	-	-	-	-	-	-	-
DINOv2 [23]	-	-	86.7	-	-	-	-	-	-	-	-	-	-
VILA-U [35]	1.80	73.3	-	-	-	-	-	-	-	-	-	-	-
UniTok [21]	0.41	70.8	-	1.4B	-	2.51	216.7	0.82	0.57	2.77	227.5	0.81	0.57
Convergence Efficiency													
REPA [40]	0.61	-	-	675M	80	7.90	122.6	0.70	0.65	-	-	-	-
DDT [34]	0.61	-	-	675M	80	6.62	135.2	0.69	**0.67**	1.52	263.7	0.78	0.63
VA-VAE [39]	**0.28**	-	-	675M	80	4.29	-	-	-	-	-	-	-
REPA-E [19]	**0.28**	-	-	675M	80	3.46	159.8	0.77	0.63	1.67	**266.3**	**0.80**	**0.63**
RAE [44]	0.57	-	84.5	675M	80	4.28	-	-	-	-	-	-	-
RAE [44]	0.57	-	84.5	835M	80	2.16	214.8	0.82	0.59	-	-	-	-
VTP (Ours)	0.36	**78.2**	**85.7**	675M	80	**2.62**	**197.8**	**0.79**	0.62	**1.44**	238.2	**0.80**	**0.63**
VTP (Ours)	0.36	78.2	85.7	1.0B	80	2.03	219.4	0.80	0.62	-	-	-	-
Long Period Training													
DiT [24]	-	-	-	675M	1400	9.62	121.5	0.67	0.67	2.27	278.2	**0.83**	0.57
SiT [22]	-	-	-	675M	1400	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
VA-VAE [39]	**0.28**	-	-	675M	800	2.17	205.6	0.77	0.65	1.35	295.3	0.79	0.65
REPA [40]	0.61	-	-	675M	800	5.78	158.3	0.70	0.68	1.29	306.3	0.79	0.64
DDT [34]	0.61	-	-	675M	400	6.27	154.7	0.68	**0.69**	1.26	**310.6**	0.79	0.65
REPA-E [19]	**0.28**	-	-	675M	800	**1.70**	217.3	0.77	0.66	1.15	304.0	0.79	0.66
RAE [44]	0.57	-	84.5	676M	800	1.87	209.7	**0.80**	0.63	1.41	309.4	0.80	0.63
RAE* [44]	0.57	-	84.5	839M	800	1.51	242.9	0.79	0.63	1.13	262.6	0.78	0.67
VTP (Ours)	0.36	**78.2**	**85.7**	675M	600	1.85	**232.3**	0.79	0.63	**1.11**	279.5	0.79	**0.67**

Table 4: Generation Performance on ImageNet 256×256.

settings outperform the AE-only baseline. Moreover, increasing the CLIP batch size consistently improves downstream generation, reducing FID from 31.89 to 27.83. This suggests that perception-oriented scaling benefits generation.

7 Conclusion

In this work, we redesign a scalable paradigm for visual tokenizer pre-training that substantially enhances generative performance and propose VTP. We demonstrate that perception-oriented tokenizer pre-training unlocks a new scaling law for generation: (1) semantic understanding is the key driver for improving generation, and (2) with this understanding, the visual tokenizer achieves scalable performance on generative tasks. In contrast to traditional tokenizers pre-trained solely on reconstruction—whose performance saturates with a small scale—our approach consistently attains a significant gain in generative performance when scaling the compute budget, model size, data scales. We hope VTP will inspire future research on visual tokenizers.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (NSFC U25B2067).

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
3. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. arXiv preprint arXiv:2502.02492 (2025)
4. Chen, B., Bi, S., Tan, H., Zhang, H., Zhang, T., Li, Z., Xiong, Y., Zhang, J., Zhang, K.: Aligning visual foundation encoders to tokenizers for diffusion models. In: The Fourteenth International Conference on Learning Representations (2025)
5. Chen, H., Han, Y., Chen, F., Li, X., Wang, Y., Wang, J., Wang, Z., Liu, Z., Zou, D., Raj, B.: Masked autoencoders are effective tokenizers for diffusion models. In: Forty-second International Conference on Machine Learning (2025)
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
10. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
11. Hansen-Estruch, P., Yan, D., Chung, C.Y., Zohar, O., Wang, J., Hou, T., Xu, T., Vishwanath, S., Vajda, P., Chen, X.: Learnings from scaling visual tokenizers for reconstruction and generation. arXiv preprint arXiv:2501.09755 (2025)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Henry, A., Dachapally, P.R., Pawar, S.S., Chen, Y.: Query-key normalization for transformers. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4246–4253 (2020)
15. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`

16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Kouzelis, T., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Eq-vae: Equivariance regularized latent space for improved generative image modeling. arXiv preprint arXiv:2502.09509 (2025)
18. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Boosting generative image modeling via joint image-feature synthesis. arXiv preprint arXiv:2504.16064 (2025)
19. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. arXiv preprint arXiv:2504.10483 (2025)
20. Li, X., Qiu, K., Chen, H., Kuen, J., Gu, J., Raj, B., Lin, Z.: Imagefolder: Autoregressive image generation with folded tokens. arXiv preprint arXiv:2410.01756 (2024)
21. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. Advances in Neural Information Processing Systems **38**, 129274–129297 (2026)
22. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
29. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
30. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
31. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)
32. Skorokhodov, I., Girish, S., Hu, B., Menapace, W., Li, Y., Abdal, R., Tulyakov, S., Siarohin, A.: Improving the diffusability of autoencoders. arXiv preprint arXiv:2502.14831 (2025)

33. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
34. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. arXiv preprint arXiv:2504.05741 (2025)
35. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. In: International Conference on Learning Representations. vol. 2025, pp. 93620–93638 (2025)
36. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
37. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. arXiv preprint arXiv:2410.10629 (2024)
38. Yang, J., Li, T., Fan, L., Tian, Y., Wang, Y.: Latent denoising makes good visual tokenizers. arXiv preprint arXiv:2507.15856 (2025)
39. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025)
40. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
41. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
43. Zheng, A., Wen, X., Zhang, X., Ma, C., Wang, T., Yu, G., Zhang, X., Qi, X.: Vision foundation models as effective visual tokenizers for autoregressive image generation. arXiv preprint arXiv:2507.08441 (2025)
44. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
45. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)