

# Towards Scalable Pre-training of Visual Tokenizers for Generation

Jingfeng Yao<sup>1\*</sup> Yuda Song<sup>2</sup> Yucong Zhou<sup>2</sup> Xinggang Wang<sup>1†</sup>

<sup>1</sup>Huazhong University of Science and Technology

<sup>2</sup>MiniMax

## 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 foundation flaw: better pixel-level 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 large-scale study reveals two principal findings: (1) understanding is a key driver of generation, and (2) much better scaling properties, where generative performance scales effectively with compute, parameters, and data allocated to the pretraining of the visual tokenizer. After large-scale pre-training, our tokenizer delivers a competitive profile (78.2 zero-shot accuracy and 0.36 rFID on ImageNet) and  $4.1\times$  faster convergence on generation compared to advanced distillation methods. More importantly, it scales effectively: without modifying standard DiT training specs, solely investing more FLOPS in pretraining VTP achieves 65.8% FID improvement in downstream generation, while conventional autoencoder stagnates very early at 1/10 FLOPS. Our pre-trained models are available at <https://github.com/MiniMax-AI/VTP>.

## 1. Introduction

Latent Diffusion Models (LDMs) [25] employ a visual tokenizer, such as a VAE [15], to compress visual signals into a latent space. Typically, visual tokenizers are pre-trained in a separate stage using a reconstruction objective.

\*Work done during an internship at MiniMax.

†Corresponding author: xgwang@hust.edu.cn.

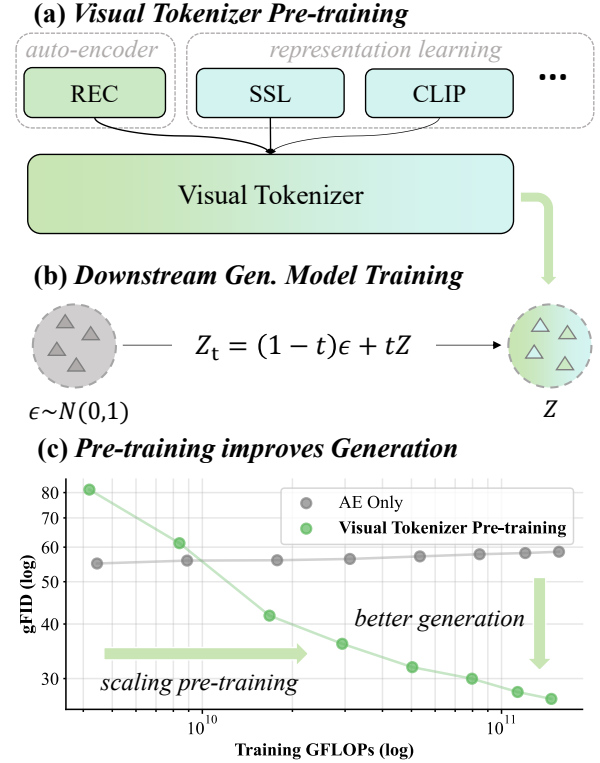


Figure 1. **Visual Tokenizer Pre-training.** We revisit the visual tokenizer pre-training in LDM [25] from a representation learning perspective. Critically, while keeping the diffusion model (e.g., DiT [21]) training configuration fixed, our method improves generation solely by scaling the tokenizer’s 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, 31, 33]. It implies that scaling the computational investment in pre-training, while potentially further improving reconstruction performance, carries the risk of compromising generation performance (see Figure 1 (c)), which is consistent with prior work [10]. We note that this limitation of reconstruction-only training may arise be-

cause 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 [3] or leveraging powerful pre-trained features [16] to the latent space, like VideoJAM and ReDi. Others implicitly structure the space using semantic constraints, an approach seen in methods like VA-VAE [33] and REPA-E [17], 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) [7] 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 [14, 23]. This is complemented by integrating established self-supervised learning techniques, notably self-distillation and mask image modeling [2, 12, 20, 37], 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 establish two principal findings: (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

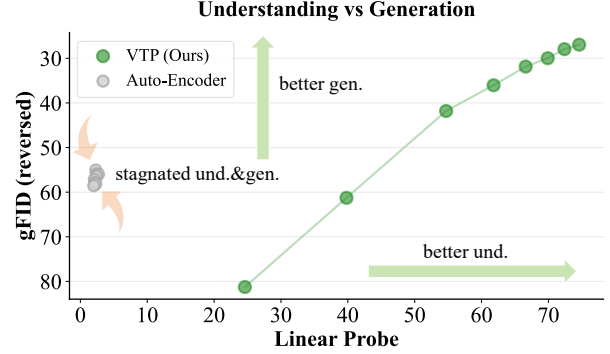


Figure 2. **Understanding is a key driver of generation.** We observe a strong positive correlation between the comprehension and generative capabilities of the latent space during visual tokenizer pre-training.

meaningful representations, leading to significant gains in downstream generation. (see Figure 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. (Sec. 4)

Our final model achieves 78.2% zero-shot accuracy and 0.36 rFID on ImageNet. Compared to previous latent space optimization methods based on distilling foundation models [33], VTP is redesigned from the ground up at the pre-training stage, achieving a superior performance ceiling and a 4.1× faster convergence speed. More importantly, our method is highly scalable. While the baseline AE hits a performance plateau at a very small scale, our method continues to improve even when the computational budget is scaled beyond 10× (see Figure 7), and achieves significant performance gains of 65.8%. This conclusively demonstrates the substantial potential of our approach. (Sec. 5)

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

- We formulate the visual tokenizer pre-training scaling problem and propose VTP, a redesigned tokenizer pre-training paradigm that explicitly optimizes the latent space for downstream generation, without modifying the downstream generative training paradigm.
- We establish two key findings: (1) there is a strong positive correlation between the understanding and generation capabilities of the latent space; and (2) VTP is the first tokenizer to demonstrate compelling scalability for generation with respect to the computation, model size, and data scale.
- VTP achieves a highly competitive profile with 78.2%

zero-shot accuracy and 0.36 rFID and achieves  $4.1\times$  faster generation convergence than prior methods on ImageNet.

## 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 [7, 11]—and transferred weights to downstream tasks like detection [24] and segmentation [30]. A recent paradigm shift has focused on weakly-supervised and unsupervised methods to enable pretraining at larger scales. For instance, CLIP [23] uses image-text contrastive learning with weak supervision by minimizing the distance between image and text features. SigLIP [34] 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 [37] and DINOv2 [20] combine MIM with self-distillation to learn more generalized representations.

Despite these advances, within the explicitly decoupled, two-stage framework of LDMs [25]—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. Latents with Pre-trained Representations

Previous 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 [33] aligns its latent space with features of visual foundation models to alleviate the trade-off between reconstruction and generation. ImageFolder [18] decouples semantic and pixel-level feature spaces to improve autoregressive generation. MAETok [4] enhances latent representations by incorporating DINOv2 features into its MIM pre-training objective. REPA-E [17] concurrently optimizes the feature space of the VAE during DiT training by leveraging supervision from a pre-trained foundation model. I-DeTok [32] 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. For instance, BLIP3-o [5] regresses SigLIP features and employs an SD-XL-based [22] decoder to boost efficiency. Recently, RAE [36] leverages DINOv2 features and trains a separate pixel decoder for reconstruction.

However, these 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 Figure 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 [8] 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  $\mathcal{L}_{\text{perceptual}}$  [35] 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:

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

### 3.3. Self-Supervised Learning

Following DINOv2 [20], our self-supervised learning framework comprises two components: masked image modeling (MIM) [12, 37] and self-distillation [2].

For a given image  $I$ , we apply data augmentation to obtain global and local views  $I_{\text{global}}$  and  $I_{\text{local}}$ . In MIM, adhering to [37],  $I_{\text{global}}$  is patch-embedded and fed directly to an EMA teacher, while its masked version is processed by

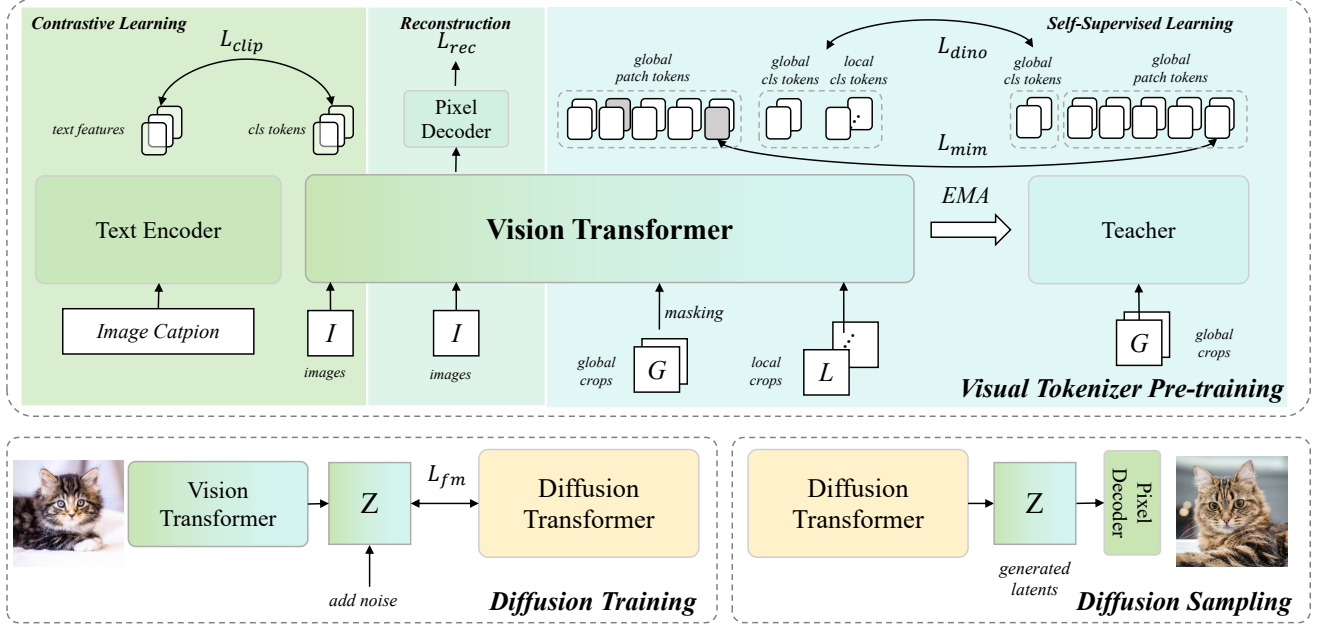


Figure 3. **Overview of Visual Tokenizer Pre-training (VTP).** By integrating representation learning (image-text contrastive [23] and self-supervised learning [20]) with reconstruction within a Vision Transformer Auto-Encoder, we find that VTP exhibits a **well-behaved scaling property** for generative performance.

the visual tokenizer, optimizing the complementary masking loss  $\mathcal{L}_{\text{mim}}$ . For self-distillation, similar to [2],  $I_{\text{global}}$  and  $I_{\text{local}}$  are passed to the visual tokenizer, and  $I_{\text{global}}$  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:

$$\mathcal{L}_{\text{ssl}} = \mathcal{L}_{\text{mim}} + \mathcal{L}_{\text{dino}} \quad (1)$$

### 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:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{ssl}} \mathcal{L}_{\text{ssl}} + \lambda_{\text{clip}} \mathcal{L}_{\text{clip}} \quad (2)$$

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.  $B_{\text{clip}} = B$ .  $B_{\text{ssl}}$  and  $B_{\text{rec}}$  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 [26]. We incorporate QKNorm [13] to enhance training stability. We employ a 12-layer transformer with a hidden dimension of 768 as



Arch.	FLOPs	#Params	rPSNR	gFID
CNN [25]	389.4G	70.3M	30.63	59.53
ViT-B [7]	87.7G	171.2M	30.72	58.40
ViT-L [7]	311.1G	607.2M	<b>31.28</b>	<b>53.51</b>

Table 1. **AutoEncoder Performance with Different Architectures.** ViT delivers better performance and efficiency under identical specifications.

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 [19], to balance semantic comprehension with reconstruction quality. An ablation study on this configuration is conducted by varying the dimension to 256. We use an internally filtered version of DataComp-1B [6] with 277M samples for tokenizer pretraining, and ImageNet [6] 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  $\lambda_{\text{clip}}$  and  $\lambda_{\text{ssl}}$  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 [20] and OpenCLIP [14].

**Downstream DiT training & evaluation** We train the Diffusion Transformer (DiT) [21] under a fixed configuration to evaluate the generative capability of our visual tokenizer. Specifically, we follow LightningDiT [33] as a strong baseline. We report FID-10k scores obtained with a LightningDiT-B [33] model trained on ImageNet [6] 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 [20, 26] 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

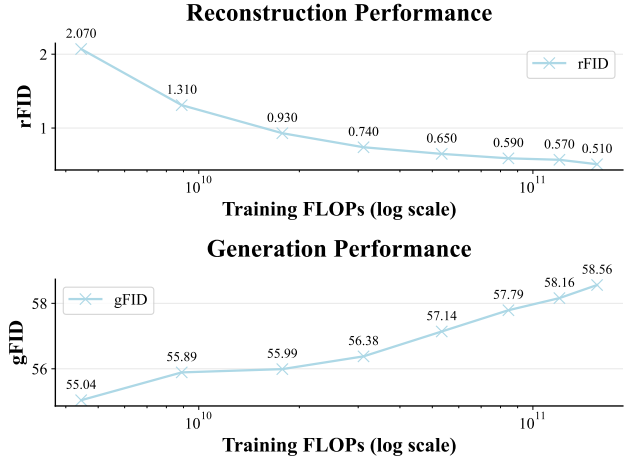


Figure 4. **Scaling up Visual Tokenizer Training with Reconstruction Only.** As training progresses, the tokenizer’s reconstruction performance improves, while its generative performance degrades concurrently. It reveals the *inadequacy of pure reconstruction tasks for scalable tokenizer pre-training*.

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 [25] under the same specifications. We use ImageNet at 256 resolution for training and testing.

As illustrated in Table 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 [10, 27], 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 from 2.0 to 0.5. However,

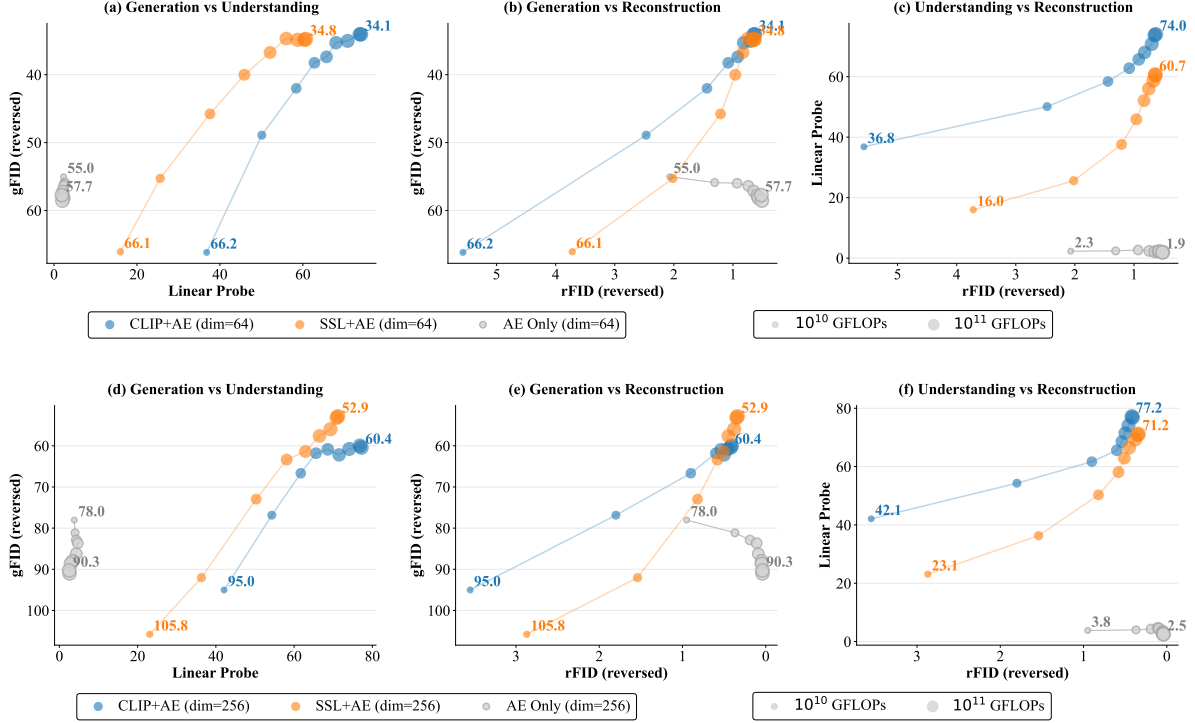


Figure 5. **Scalability of CLIP+AE & SSL+AE Visual Tokenizer Pre-training.** Scaling properties under different strategies and bottleneck dimensions. Our method shows correlated growth in generation and comprehension with compute, while VAE-based tokenizer performance rapidly saturates.

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) [23] or self-supervised learning (SSL, specifically DINOv2) [20] 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 train-

ing 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 Figure 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 Figure 5 (b)&(e) illustrates a negative yield in reconstruction-only AEs: as the computational budget increases, reconstruction performance improves but the generation performance degrades.

(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 Figure 5 (a)&(d) and (b)&(e). Specifically, our visual tokenizer pre-training, which jointly optimizes for reconstruction and representation learning, enables continuous, concurrent im-

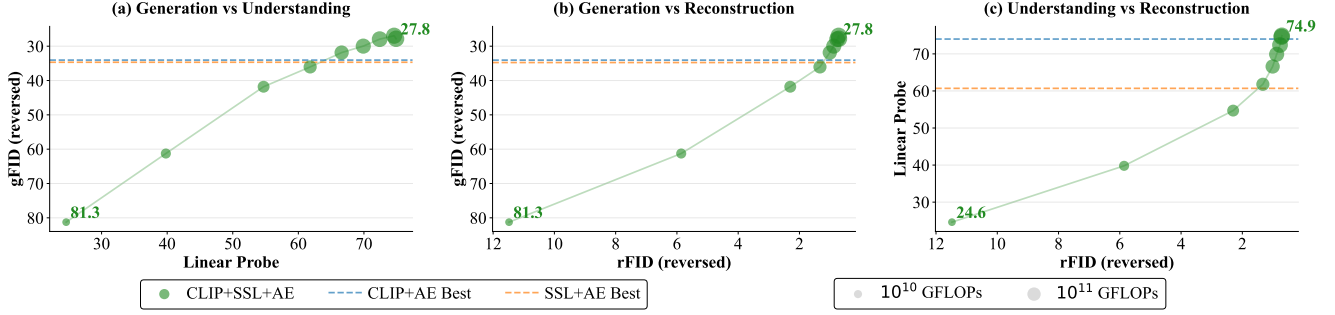


Figure 6. **Scalability of CLIP+SSL+AE Visual Tokenizer Pre-training.** Under the same computational budget, the f16d64 tokenizer trained with joint CLIP and SSL representation learning achieves the best performance in both generation and comprehension.

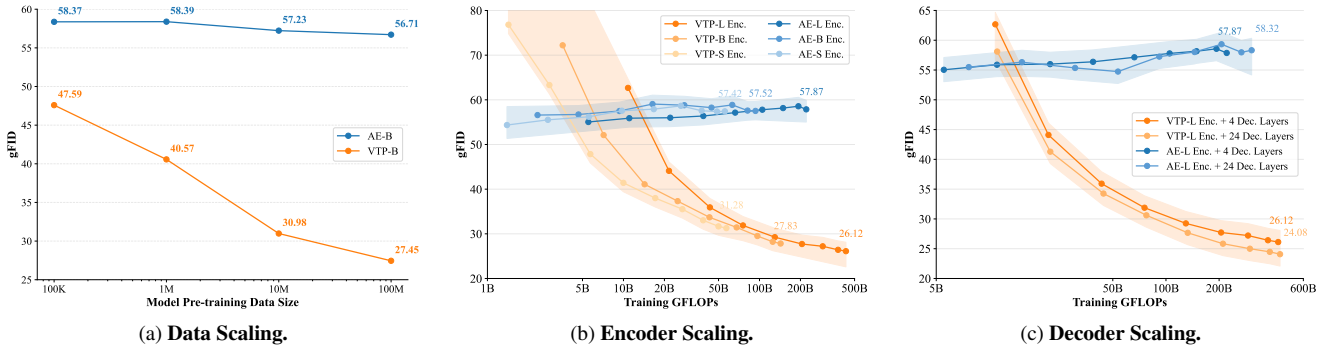


Figure 7. **Scalability of data and parameters.** We observe a new scaling property: the DiT generation performance within fixed training FLOPs increases while its tokenizer has larger model sizes and training data.

provement 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 Figure 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 pro-

Model	Und.		Rec.	Gen.
	Zero-shot	Lin. Prob.	rFID	80ep w/o cfg
<i>Discriminative Baselines</i>				
OpenCLIP [14]	74.0	-	-	-
CLIP [23]	75.5	-	-	-
SigLIP [34]	<b>80.5</b>	-	-	-
MAE [12]	-	85.9	-	-
DINOv2 [20]	-	<b>86.7</b>	-	-
<i>Visual Tokenizers</i>				
SD-VAE [25]	-	-	0.63	7.13
UniTok [19]	70.8*	-	0.41	-
VILA-U [29]	73.3	-	1.80	-
VA-VAE-d32 [33]	-	-	0.28	4.29
VA-VAE-d64 [33]	-	-	<b>0.15</b>	-
RAE-d768 [36]	-	84.5	0.57	4.28 <sup>†</sup>
<b>VTP-S-d64</b>	66.7	77.5	0.98	5.46
<b>VTP-B-d64</b>	73.2	81.0	0.74	3.88
<b>VTP-L-d64</b>	<b>78.2</b>	<b>85.7</b>	0.36	<b>2.81</b>

Note: 1) All results are cited from original papers. 2) \*We take UniTok's results trained from scratch for fair comparison. 3) <sup>†</sup>We take RAE's results trained with the same generation model for fair comparison.

Table 2. **Comprehensive Comparison.** We evaluate Understanding (Zero-shot Acc. & Linear Probing), Reconstruction (rFID), and Generation capabilities. For Generation, we report FID-50K scores using LightningDiT trained on ImageNet for 80 epochs and tested without CFG. Gray values indicate specialized baselines.

ceed 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.



Figure 8. **Reconstruction Comparison.** VTP demonstrates superior reconstruction performance, particularly in terms of color accuracy and the preservation of fine textures.

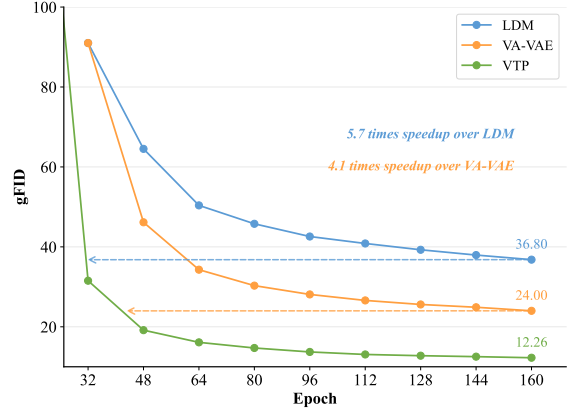


Figure 9. **Generation Comparison.** VTP achieves faster convergence in generation, indicating a higher potential upper bound for the pre-training paradigm.

## 5. Further scaling and Comparison

Given the prevalence of foundation models [20, 23, 26, 28, 34], a critical question arises: why is it necessary to explore re-training and new scaling strategies? To answer this, we scale up three tokenizers with symmetric ViT architectures using  $3\times$  training compute, denoted as VTP-S, VTP-B, and VTP-L. The results are presented in Table 2.

There are two representative approaches to utilizing representations for generation: distillation-based methods [33] and methods based on fixed representation encoders [36]. Regarding the former, we observe that distillation loss fails to fully leverage the performance of understanding models. As shown in Figure 9, although VA-VAE achieves significantly faster convergence compared to LDM, its performance ceiling is notably lower than that of our VTP. Specifically, VTP improves convergence speed by  $4.1\times$ . Regard-



ing the latter, as illustrated in Figure 8, we find that reconstructing from fixed encoder representations leads to distinct issues such as color shifts and texture errors. These artifacts negatively impact the quality of generated content. In contrast, training from scratch effectively integrates understanding and reconstruction representations, thereby significantly mitigating these issues. Consequently, the exploration of scaling laws in VTP demonstrates greater potential for both reconstruction and generation tasks.

## 6. Conclusion and Future Work

In this work, we redesign a scalable paradigm for visual tokenizer pre-training that substantially enhances generative performance. We demonstrate that (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 stark contrast to traditional tokenizers pre-trained solely on reconstruction—whose performance saturates with a small scale—our approach attains a significant 65.8% relative gain in generative performance when scaling the compute budget by 10 $\times$ .

**Future Work** This work opens up two new research avenues. First, we provide the first evidence that representation learning objectives like CLIP and SSL are key drivers for both generative quality and pre-training scalability in visual tokenizers. This naturally raises the question: what other perceptual tasks could be integrated to form even more powerful multi-objective pre-training frameworks? Second, we demonstrate for the first time the critical importance of data scaling for the tokenizer itself. This finding suggests that not just the amount, but also the distribution of pre-training data could be a powerful lever. For example, tokenizers pre-trained on data rich with specific attributes (e.g., text rendering) could unlock specialized capabilities for corresponding downstream generative tasks.

## References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3, 4
- [3] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025. 2
- [4] Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. 3
- [5] Jiahui Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 3
- [6] 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*, pages 248–255. Ieee, 2009. 5
- [7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 5
- [8] 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*, pages 12873–12883, 2021. 3
- [9] 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. 1
- [10] Philippe Hansen-Estruch, David Yan, Ching-Yao Chung, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from scaling visual tokenizers for reconstruction and generation. *arXiv preprint arXiv:2501.09755*, 2025. 1, 5
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 8
- [13] Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245*, 2020. 4
- [14] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. Computer software. An open-source implementation of CLIP. 2, 5, 8
- [15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1
- [16] Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. *arXiv preprint arXiv:2504.16064*, 2025. 2
- [17] 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. 2, 3

- [18] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024. 3
- [19] 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. 5, 8
- [20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4, 5, 6, 8
- [21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 5
- [22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3, 4, 6, 8
- [24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3
- [25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 5, 8
- [26] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4, 5, 8
- [27] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 5
- [28] Michael Tschanen, 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. 8
- [29] 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. 8
- [30] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3
- [31] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 1
- [32] Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers. *arXiv preprint arXiv:2507.15856*, 2025. 3
- [33] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025. 1, 2, 3, 5, 8
- [34] 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*, pages 11975–11986, 2023. 3, 8
- [35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3
- [36] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 3, 8
- [37] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3