Reconstruction *v.s.* Generation: Taming Optimization Dilemma in Latent Diffusion Models

Jingfeng Yao, Xinggang Wang*
Huazhong University of Science and Technology

Abstract

*Latent diffusion models (LDMs) with Transformer architectures excel at generating high-fidelity images. However, recent studies reveal an **optimization dilemma** in this two-stage design: while increasing the per-token feature dimension in visual tokenizers improves reconstruction quality, it requires substantially larger diffusion models and more training iterations to achieve comparable generation performance. Consequently, existing systems often settle for sub-optimal solutions, either producing visual artifacts due to information loss within tokenizers or failing to converge fully due to expensive computation costs. We argue that this dilemma stems from the inherent difficulty in learning unconstrained high-dimensional latent spaces. To address this, we propose aligning the latent space with pre-trained vision foundation models when training the visual tokenizers. Our proposed **VA-VAE** (Vision foundation model Aligned Variational AutoEncoder) significantly expands the reconstruction-generation frontier of latent diffusion models, enabling faster convergence of Diffusion Transformers (DiT) in high-dimensional latent spaces. To exploit the full potential of VA-VAE, we build an enhanced DiT baseline with improved training strategies and architecture designs, termed **LightningDiT**. The integrated system demonstrates remarkable training efficiency by reaching FID=2.11 in just 64 epochs – an over 21× convergence speedup over the original DiT implementations, while achieving state-of-the-art performance on ImageNet-256 image generation with FID=1.35. Models and codes are available at https://github.com/hustvl/LightningDiT.*

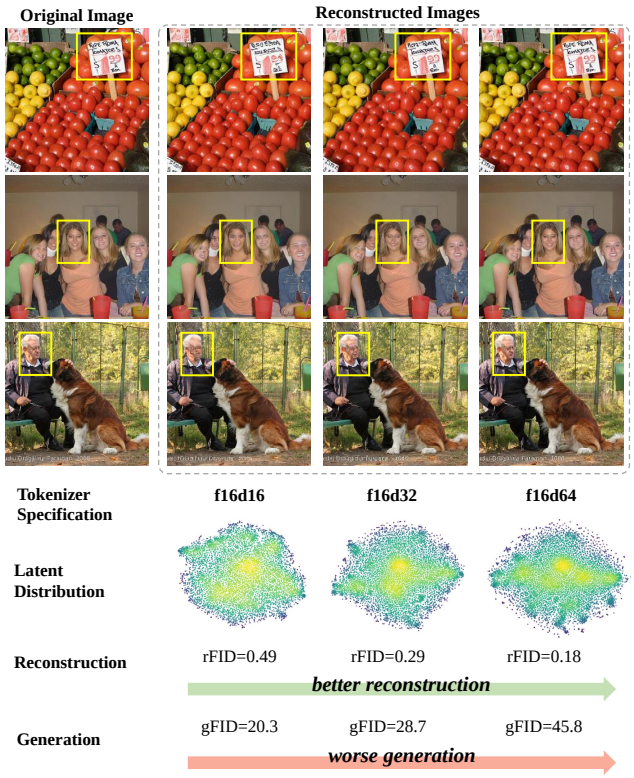

Figure 1. **Optimization dilemma within latent diffusion models.** In latent diffusion models, increasing the dimension of the visual tokenizer enhances detail reconstruction but significantly reduces generation quality. (In tokenizer specification, "f" and "d" represent the downsampling rate and dimension, respectively. All results are evaluated on ImageNet 256×256 dataset with a fixed compute budget during diffusion model training.)

1. Introduction

The latent diffusion model [33] utilizes a continuous-valued variational autoencoder (VAE) [17], or visual tokenizer, to compress visual signals and thereby reduce the computational demands of high-resolution image generation. The performance of these visual tokenizers, particularly their compression and reconstruction capabilities, plays a crucial role in determining the overall system effectiveness [5, 7]. A straightforward approach to enhance the reconstruction capability is to increase the feature dimension of visual tokens, which effectively expands the information capacity of the latent representation. Recently, several influential text-

*Corresponding author: xgwang@hust.edu.cn

to-image works [5, 7, 20] have explored higher-dimensional tokenizers compared to the widely adopted VAE in Stable Diffusion [30, 33], as these tokenizers offer improved detail reconstruction, enabling finer generative quality.

However, as research has advanced, an *optimization dilemma* has emerged between reconstruction and generation performance in latent diffusion models [7, 13, 16]. Specifically, while increasing token feature dimension improves the tokenizer's reconstruction accuracy, it significantly degrades the generation performance (see Fig. 1). Currently, two common strategies exist to address this issue: the first involves scaling up model parameters, as demonstrated by Stable Diffusion 3 [7], which shows that higher-dimensional tokenizers can achieve stronger generation performance with a significantly larger model capacity—however, this approach requires significantly more training compute, making it prohibitively expensive for most practical applications. The second strategy is to deliberately limit the tokenizer's reconstruction capacity,e.g. DC-AE or W.A.L.T. [4, 13], for faster convergence of diffusion model training. Yet, this compromised reconstruction quality inherently limits the upper bound of generation performance, leading to imperfect visual details in generated results. Both approaches involve inherent trade-offs and do not fundamentally resolve the underlying optimization dilemma.

In this paper, we propose a straightforward yet effective approach to address this optimization dilemma. We draw inspiration from Auto-Regressive (AR) generation, where increasing the codebook size of discrete-valued VAEs leads to low codebook utilization [42, 47]. Through visualizing the latent space distributions across different feature dimensions (see Fig. 1), we observe that higher-dimensional tokenizers learn latent representations in a less spread-out manner, evidenced by more concentrated high-intensity areas in the distribution visualization. This analysis suggests that the optimization dilemma stems from the inherent difficulty of learning unconstrained high-dimensional latent spaces from scratch. To address this issue, we develop a vision foundation model guided optimization strategy for continuous VAEs [17] in latent diffusion models. Our key finding demonstrates that learning latent representations guided by vision foundation models significantly enhances the generation performance of high-dimensional tokenizers while preserving their original reconstruction capabilities (see Fig. 2).

Our main technical contribution is the Vision Foundation model alignment Loss (**VF Loss**), a plug-and-play module that aligns latent representations with pre-trained vision foundation models [15, 28] during tokenizer training. While naively initializing VAE encoders with pre-trained vision foundation models has proven ineffective [22]—likely because the latent representation quickly deviates from its initial state to optimize reconstruction—we find that a carefully designed joint reconstruction and alignment loss is crucial. Our alignment loss is specifically crafted to regularize high-dimensional latent spaces without overly constraining their capacity. First, we enforce both element-wise and pair-wise similarities to ensure comprehensive regularization of global and local structures in the feature space. Second, we introduce a margin in the similarity cost to provide controlled flexibility in the alignment, thereby preventing over-regularization. Additionally, we investigate the impact of different vision foundation models.

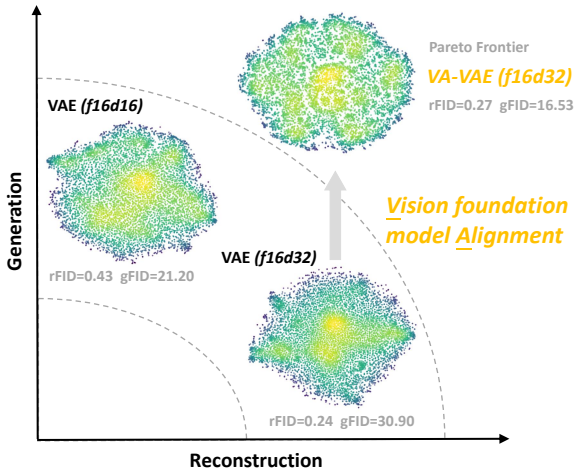

Figure 2. **Reconstruction-generation frontier of latent diffusion models.** VA-VAE improves the feature distribution of high-dimensional latent. Through alignment with vision foundation models, we expand the frontier between reconstruction and generation in latent diffusion models.

To evaluate the generation performance, we couple the proposed Vision foundation model Aligned VAE (**VA-VAE**) with Diffusion Transformers (DiT) [29] to create a latent diffusion model. To fully exploit the potential of VA-VAE, we build an enhanced DiT framework through advanced diffusion training strategies and transformer architectural improvements, which we name **LightningDiT**. Our contributions achieve the following significant milestones:

- The proposed VF Loss effectively resolves the optimization dilemma in latent diffusion models, enabling over 2.5× faster DiT training with high-dimensional tokenizers;
- The integrated system reaches an FID of 2.11 within just 64 training epochs, an over 21× convergence speedup compared with the original DiT;
- The integrated system achieves a state-of-the-art FID score of 1.35 on ImageNet-256 image generation.

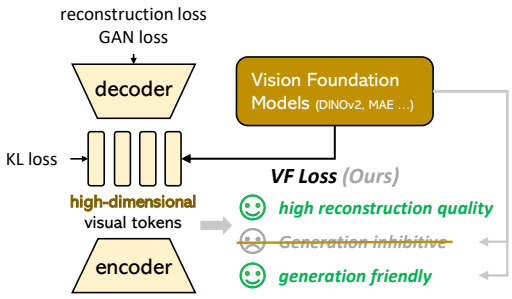

Figure 3. **The proposed Vision foundation model Aligned VAE (VA-VAE).** Vision foundation models are used to guide the training of high-dimensional visual tokenizers, effectively mitigating the optimization dilemma and improve generation performance.

2. Related Work

2.1. Tokenizers for Visual Generation

Visual tokenizers, represented by variational autoencoders (VAEs) [17], leverage an encoder-decoder architecture to create continuous visual representations, facilitating the compression and reconstruction of visual signals. While VAEs operate in continuous space, VQVAE [39] introduces discrete representation learning through a learnable codebook for quantization. VQGAN [6] further enhances this discrete approach by incorporating adversarial training, establishing itself as a cornerstone for autoregressive (AR) generation frameworks [2]. However, these discrete approaches face a fundamental challenge: as revealed in [42], larger codebooks improve reconstruction fidelity but lead to poor codebook utilization, adversely affecting generation performance. Recent works propose different solutions to this utilization problem: MAGVIT-v2 [42] introduces Look-up Free Quantization (LFQ), while VQGAN-LC [47] leverages CLIP [32] vision features for codebook initialization, achieving near-complete codebook utilization.

Interestingly, continuous VAE-based latent diffusion systems [33] encounter a parallel optimization challenge: increasing the tokenizer's feature dimensionality enhances reconstruction quality but degrades generation performance, necessitating substantially higher training costs [4, 7, 22]. Despite the significance of this trade-off in both discrete and continuous domains, current literature lacks comprehensive analysis and effective solutions for continuous VAE optimization. Our work addresses this fundamental limitation by introducing vision foundation model alignment into continuous VAE training. This principled approach resolves the optimization dilemma by structuring the latent space according to well-established visual representations, enabling efficient training of diffusion models in higher-dimensional latent spaces. Importantly, our solution maintains the computational efficiency of diffusion model training as it requires no additional parameters, while significantly accelerating convergence by over 2.5 $\times$.

2.2. Fast Convergence of Diffusion Transformers

Diffusion Transformers [29] are currently the most popular implementation of latent diffusion models [3, 7, 10, 25, 27]. Due to their remarkable scalability, they have proven effective across various text-to-image and text-to-video tasks. However, they suffer from slow convergence speeds. Previous works have proposed various solutions: SiT [26] enhances DiT's efficiency through Rectified Flow integration, while MDT [11] and MaskDiT [45] achieve faster convergence by incorporating mask image modeling. REPA [43] accelerates convergence by aligning DiT features with vision foundation models like DINOv2 [28] during training.

Different from these approaches, we identify that a major bottleneck in efficient latent diffusion training lies in the tokenizer itself. We propose a principled approach that optimizes the latent space learned by the visual tokenizer. Unlike methods that combine diffusion loss with auxiliary losses [11, 43, 45], which incur additional computational costs during training, our approach achieves faster convergence without modifying the diffusion models. Additionally, we add several optimizations in terms of training strategies and architecture designs to the original DiT implementation that further accelerate training.

[*]*Relationship to REPA [43]*: Both our work and REPA [43] utilize vision foundation models to aid in diffusion training, yet the motivations and approaches are entirely different. REPA aims to employ vision foundation models to constrain DiT, thereby *enhancing the convergence speed* of generative models. In contrast, our work takes into account both the reconstruction and generative capabilities within the latent diffusion model, with the objective of leveraging foundation models to regulate the high-dimensional latent space of the tokenizer, thereby *resolving the optimization conflict* between the tokenizer and the generative model.

3. Align VAE with Vision Foundation Models

In this section, we introduce VA-VAE, a visual tokenizer trained through vision foundation model alignment. The key approach involves constraining the tokenizer's latent space by leveraging the feature space of the foundation model, which enhances its suitability for generative tasks.

As illustrated in Figure 3, our architecture and training process mainly follows LDM [33], employing a VQGAN [6] model architecture with a continuous latent space, constrained by KL loss. Our key contribution lies in the design of **V**ision **F**oundation model alignment loss, **VF loss**, which substantially optimizes the latent space without altering the model architecture or training pipeline, effectively resolving the optimization dilemma discussed in Section 1.

The VF loss consists of two components: marginal cosine similarity loss and marginal distance matrix similarity loss. These components are carefully crafted as a straightforward, plug-and-play module that is decoupled from the VAE architecture. We will explain each part in detail below.

3.1. Marginal Cosine Similarity Loss

During training, a given image I is processed by both the encoder of visual tokenizer and a frozen vision foundation model, resulting in image latents Z and foundational visual representations F. As shown in Eq. 1, we project Z to match the dimensionality of F using a linear transformation, where $W \in \mathbb{R}^{d_f \times d_z}$ respectively, producing $Z' \in \mathbb{R}^{d_f}$.

$$Z' = WZ \tag{1}$$

As defined in Eq. 2, the loss function $\mathcal{L}_{\text{mcos}}$ minimizes the similarity gap between corresponding features z'_{ij} and f_{ij} from feature matrices Z' and F at each spatial location (i, j). For each pair, we compute the cosine similarity $\frac{z'_{ij} \cdot f_{ij}}{\|z'_{ij}\|\|f_{ij}\|}$ and subtract a margin m_1. The ReLU function ensures that only pairs with similarities below m_1 contribute to the loss, focusing alignment on less similar pairs. The final loss is averaged over all positions in the $h \times w$ feature grid.

$$\mathcal{L}_{\text{mcos}} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \text{ReLU}\left(1 - m_1 - \frac{z'_{ij} \cdot f_{ij}}{\|z'_{ij}\|\|f_{ij}\|}\right) \tag{2}$$

3.2. Marginal Distance Matrix Similarity Loss

Complementary to $\mathcal{L}_{\text{mcos}}$, which enforces point-to-point absolute alignment, we also aim for the relative distribution distance matrices within the features to be as similar as possible. To achieve this, we propose the marginal distance matrix similarity loss.

In Eq. 3, the distance matrix similarity Loss aligns the internal distributions of feature matrices z and f. Here, $N = h \times w$ represents the total number of elements in each flattened feature map. For each pair (i, j), we compute the absolute value of the cosine similarity difference between the corresponding vectors in feature matrices z and f, thus promoting closer alignment of their relative structures. Similarly, we subtract a margin m_2 to relax the constraint. The ReLU function ensures that only pairs with differences exceeding m_2 contribute to the loss.

$$\mathcal{L}_{\text{mdms}} = \frac{1}{N^2} \sum_{i,j} \text{ReLU}\left(\left|\frac{z_i \cdot z_j}{\|z_i\|\|z_j\|} - \frac{f_i \cdot f_j}{\|f_i\|\|f_j\|}\right| - m_2\right) \tag{3}$$

Training Trick	Training Sample	Epoch	FID-50k↓
DiT-XL/2 [29]	400k × 256	80	19.50
Training Strategies			
+ Rectified Flow [23]	100k × 1024	80	17.20
+ *batchsize*×4 & *lr*×2			16.59
+ AdamW $\beta_2 = 0.95$ [1]			16.61
+ Logit Normal Sampling [7]			13.99
+ Velocity Direction Loss [41]			12.52
Architecture Improvements			
+ SwiGLU FFN [34]	100k × 1024	80	10.10
+ RMS Norm [44]			9.25
+ Rotary Pos Embed [35]			7.13
+ patch size=1 & VA-VAE (Sec. 3)			4.29

Table 1. **Performance of LightningDiT**. With SD-VAE [33], LightningDiT achieves FID-50k=7.13 on ImageNet class-conditional generation, using 94% fewer training samples compared to the original DiT [29]. We show that the original DiT can also achieve exceptional performance by leveraging advanced design techniques.

3.3. Adaptive Weighting

In Figure 3, the original reconstruction loss and KL loss are both sum losses, which places the VF loss on a completely different scale, making it challenging to adjust the weight for stable training. Inspired by GAN Loss [6], we employ an adaptive weighting mechanism. Before backpropagation, we calculate the gradients of L_{vf} and L_{rec} on the last convolutional layer of the encoder, as shown in Eq 4. The adaptive weighting is set as the ratio of these two gradients, ensuring that L_{vf} and L_{rec} have similar impacts on model optimization. This alignment significantly reduces the adjustment range of the VF Loss.

$$w_{\text{adaptive}} = \frac{\|\nabla L_{\text{rec}}\|}{\|\nabla L_{\text{vf}}\|} \tag{4}$$

Then we get VF Loss with adaptive weighting like Eq 5. The purpose of adaptive weighting is to quickly align our loss scales across different VAE training pipelines. On this basis, we can still use manually tuned hyperparameters to further improve performance.

We will evaluate the significant role of VF loss in achieving the latent diffusion Pareto frontier for both reconstruction and generation in our forthcoming experiments.

$$\mathcal{L}_{\text{vf}} = w_{\text{hyper}} * w_{\text{adaptive}}(\mathcal{L}_{\text{mcos}} + \mathcal{L}_{\text{mdms}}) \tag{5}$$

4. Improved Diffusion Transformer

In this section, we introduce our LightningDiT. Diffusion Transformers (DiT) [29] has gained considerable success as a foundational model for text-to-image [3, 7] and text-to-video generation tasks [27, 31]. However, its convergence speed on ImageNet is significantly slow, resulting in high experimental iteration costs. Previous influential work like

Tokenizer	Spec.	Reconstruction Performance				Generation Performance (FID-10K)↓		
		rFID↓	PSNR↑	LPIPS↓	SSIM↑	LightningDiT-B	LightningDiT-L	LightningDiT-XL
LDM [33]		0.49	26.10	0.132	0.72	16.24	9.49	8.28
LDM+VF loss (*MAE*) [15]		0.51	26.01	0.137	0.71	16.86 *(+0.62)*	10.93 *(+1.44)*	9.19 *(+0.91)*
LDM+VF loss (*DINOv2*) [28]		0.55	25.29	0.147	0.69	15.79 *(-0.45)*	10.02 *(+0.53)*	8.71 *(+0.43)*
LDM [33]		0.26	28.59	0.089	0.80	22.62	12.86	10.92
LDM+VF loss (*MAE*) [15]		0.28	28.33	0.091	0.80	19.89 *(-2.73)*	11.51 *(-1.35)*	9.92 *(-1.00)*
LDM+VF loss (*DINOv2*) [28]		0.28	27.96	0.096	0.79	15.82 *(-6.80)*	9.82 *(-3.04)*	8.22 *(-2.70)*
LDM [33]		0.17	31.03	0.055	0.88	36.83	20.73	17.24
LDM+VF loss (*MAE*) [15]		0.15	31.03	0.054	0.87	23.58 *(-13.25)*	14.40 *(-6.33)*	11.69 *(-5.55)*
LDM+VF loss (*DINOv2*) [28]		0.14	30.71	0.055	0.87	24.00 *(-12.83)*	14.95 *(-5.78)*	11.98 *(-5.26)*

Table 2. **VF loss Improves Generation Performance.** The *f16d16* tokenizer specification is widely used [21, 33]. As dimensionality increases, we observe that (1) higher dimensions improve reconstruction but reduce generation quality, highlighting an optimization dilemma within the latent diffusion framework; (2) VF Loss significantly enhances generative performance in high-dimensional tokenizers with minimal impact on reconstruction.

DINOv2 [28], ConvNeXt [24], and EVA [9] demonstrates how incorporating advanced design strategies can revitalize classic methods [14, 46]. In our work, we aim to extend the potential of the DiT architecture and explore the boundaries of how far the DiT can go. While we do not claim any individual optimization detail as our original contribution, we believe an open-source, fast-converging training pipeline for DiT will greatly support the community's ongoing research on diffusion transformers.

We utilize the SD-VAE [33] with the *f8d4* specification as the visual tokenizer and employ DiT-XL/2 as our experimental model. We show the optimization routine in Table 1. Each model has been trained for 80 epochs and sampled with a dopri5 integrator, which has less NFE than the original DiT for fast inference. To ensure a fair comparison, no sample optimization methods such as cfg interval [19] and timestep shift [20] are used. We adopt three categories of optimization strategies. At the computational level, we implement *torch.compile* [37] and *bfloat16* training for acceleration. Additionally, we increase the batch size and reduce the β_2 of AdamW to 0.95, drawing from previous work AuraFlow [8]. For diffusion optimization, we incorporate Rectified Flow [23, 26], logit normal distribution (lognorm) sampling [7], and velocity direction loss [41]. At the model architecture level, we apply common Transformer optimizations, including RMSNorm [44], SwiGLU [34], and RoPE [35]. During training, we observe that some acceleration strategies are not orthogonal. For example, gradient clipping is effective when used alone but tends to reduce performance when combined after lognorm sampling and velocity direction loss.

Our optimized model, LightningDiT, achieves an FID of 7.13 (*cfg=1*) with SD-VAE at around 80 epochs on ImageNet class-conditional generation, which is only 6% of the training volume required by the original DiT and SiT models over 1400 epochs. Previous great work MDT [11] or REPA [43] achieved similar convergence performance with the help of Mask Image Modeling (MIM) and representation alignment. Our results show that, even without any complex training pipeline, naive DiT could still achieve very competitive performance. This optimized architecture has been of great help in our following rapid experiment validation.

5. Experiments

In this section, our main objective is to achieve the reconstruction and generation frontier (see Figure 2) of reconstruction and generation within the latent diffusion system by leveraging VF loss proposed in Section 3. With the support of LightningDiT introduced in Section 4, we demonstrate how VF loss effectively resolves the optimization dilemma, from the perspective of convergence, scalability, and overall system performance.

5.1. Implementation Details

We introduce our latent diffusion system in detail. For the visual tokenizer, we employ an architecture and training strategy mainly following to LDM [33]. Specifically, we utilize the VQGAN [6] network structure, omitting quantization and applying KL Loss to regulate the continuous latent space. To enable multi-node training, we scale the learning rate and global batch size to $1e{-}4$ and 256, respectively, following a setup from MAR [21]. We train three different $f16$ tokenizers: one without VF loss, one using VF loss (*MAE*), and another using VF loss (*DINOv2*). Here f denotes the downsampling rate and d denotes the latent dimension. Empirically, we set $m_1 = 0.5$, $m_2 = 0.25$, and $w_{hyper} = 0.1$. We argue different vision foundation models might converge to different margin settings. For the generative model, we employ LightningDiT, which is

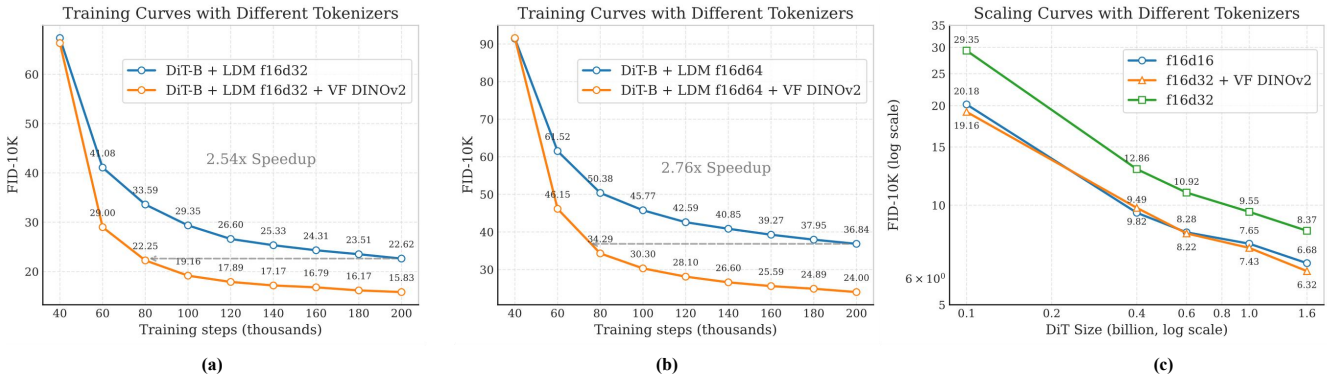

Figure 4. **(a)&(b) VF Loss Improves Convergence.** We train LightningDiT-B for 160 epochs on ImageNet at 256 resolution using different tokenizers. The VF loss significantly accelerates convergence, with a maximum speedup of up to 2.7 times. **(c) VF Loss Improves Scalability.** VF loss reduces the need for large parameters in generative models of high-dimensional tokenizer, enabling better scalability.

further refined with the design techniques outlined in Section 4. We pre-extract all latent features from the tokenizer and train various versions of LightningDiT on ImageNet at a resolution of 256 for either 80 or 160 epochs. We set the patch size of DiT to 1, ensuring that the downsampling rate of the entire system is 16. This approach is consistent with the strategy proposed in [4], i.e. all compression steps are handled by the VAE. Unless otherwise noted, our model's other architectural parameters are consistent with those of DiT [29].

5.2. Foundation Models Improve Convergence

Table 2 presents an evaluation of the reconstruction and generation of eight different tokenizers, with all generative models trained for 160 epochs (LightningDiT-B) or 80 epochs (LightningDiT-L & LightningDiT-XL) on ImageNet. Here we come with the following findings:

The results highlight the optimization dilemma in latent diffusion systems, as discussed in Section 1. The results highlighted in blue in the table illustrate the reconstruction performance (rFID) and the corresponding generation performance (FID). It can be observed that as the tokenizer dimension increases, its rFID decreases, while the corresponding generation FID increases.

The VF Loss can effectively enhance the generative performance of high-dimensional tokenizers. In the f16d32 and f16d64 sections of the table, both VF loss (*DINOv2*) and VF loss (*MAE*) significantly improve the generative performance of DiT models across different scales. This makes it possible to achieve systems with higher reconstruction performance and higher generative performance (i.e., the reconstruction-generation frontier mentioned in the introduction). It is worth noting, however, that the VF loss is unnecessary for lower-dimensional tokenizers, such as normally used f16d16 [21, 33, 36]. This stays consistent with the latent distribution observation in Figure 1. We suggest this is because lower-dimensional spaces can learn more reasonable distributions without the need for additional supervisory signals.

Additionally, we present a convergence plot of FID over training time in Figure 4 (a) & (b). On f16d32 and f16d64, the use of VF loss accelerates convergence by factors of 2.54 and 2.76, respectively. These also demonstrate that the VF loss significantly enhances the generative performance and convergence speed of high-dimensional tokenizers.

5.3. Foundation Models Improve Scalability

As discussed in Section 1, increasing the model parameter count serves as a way to improve the generative performance of high-dimensional tokenizers [7]. We use LightningDiT models ranging from 0.1B to 1.6B in size to evaluate the generative performance of 3 different tokenizers.

To facilitate the observation of the power law in scaling, we employ a log scale for the axes. We notice a slight convergence trend between the blue and green lines as the number of parameters increases, yet a significant gap remains. This implies that the negative effects on generation brought by high-dimensional f16d32 tokenizers are not fully mitigated even at 1.6B, a parameter size already considered substantial on ImageNet. We find that the VF loss effectively bridges this gap. Below 0.6B, the performance of the orange and blue lines is similar. However, as the model scales beyond 1B, f16d32 VF DINOv2 gradually distances itself from f16d16, demonstrating stronger scalability.

5.4. Convergence 21.8× Faster than DiT

We find that the VF loss (*DINOv2*) brings the most significant improvement in generative performance. Therefore, we extend the training time for the tokenizer and adopt a progressive training strategy to train the LDM VF loss (*DI-*

Method	Reconstruction		Training Epoches	#params	Generation w/o CFG					Generation w/ CFG				
	Tokenizer	rFID			gFID	sFID	IS	Pre.	Rec.	gFID	sFID	IS	Pre.	Rec.
	AutoRegressive (AR)													
MaskGIT [2]	MaskGiT	2.28	555	227M	6.18	-	182.1	0.80	0.51	-	-	-	-	-
LlamaGen [36]	VQGAN†	0.59	300	3.1B	9.38	8.24	112.9	0.69	0.67	2.18	5.97	263.3	0.81	0.58
VAR [38]	-	-	350	2.0B	-	-	-	-	-	1.80	-	365.4	0.83	0.57
MagViT-v2 [42]	-	-	1080	307M	3.65	-	200.5	-	-	1.78	-	319.4	-	-
MAR [21]	LDM†	0.53	800	945M	2.35	-	227.8	0.79	0.62	1.55	-	303.7	0.81	0.62
	Latent Diffusion Models													
MaskDiT [45]	SD-VAE [33]	0.61	1600	675M	5.69	10.34	177.90	0.74	0.60	2.28	5.67	276.6	0.80	0.61
DiT [29]			1400	675M	9.62	6.85	121.5	0.67	0.67	2.27	4.60	278.2	**0.83**	0.57
SiT [26]			1400	675M	8.61	6.32	131.7	0.68	0.67	2.06	4.50	270.3	0.82	0.59
FasterDiT [41]			400	675M	7.91	5.45	131.3	0.67	**0.69**	2.03	4.63	264.0	0.81	0.60
MDT [11]			1300	675M	6.23	5.23	143.0	0.71	0.65	1.79	4.57	283.0	0.81	0.61
MDTv2 [12]			1080	675M	-	-	-	-	-	1.58	4.52	**314.7**	0.79	0.65
REPA [43]			800	675M	5.90	-	-	-	-	1.42	4.70	305.7	0.80	0.65
LightningDiT	**VA-VAE**	**0.28**	**64**	675M	5.14	**4.22**	130.2	0.76	0.62	2.11	4.16	252.3	0.81	0.58
			800	675M	**2.17**	4.36	**205.6**	**0.77**	0.65	**1.35**	**4.15**	295.3	0.79	**0.65**

Table 3. **System-Level Performance on ImageNet 256×256.** Our latent diffusion system achieves ***state-of-the-art*** performance with rFID=0.28 and FID=1.35. Besides, our LightningDiT together with VA-VAE surpasses DiT [29] and SiT [26] in FID within only 64 training epochs, demonstrating a ***21.8 × faster convergence.***

Figure 5. **Visualization Results.** We visualize our latent diffusion system with proposed VA-VAE together with LightningDiT-XL trained on ImageNet 256 × 256 resolution.

NOv2) for 125 epochs, resulting in a VA-VAE with stronger generative capabilities through prolonged training. We train LightningDiT-XL for 800 epochs following the parameters in Table 1. Specifically, at 480 epochs, we disable the lognorm parameter to enable the near-converged network to learn more effectively across all noise intervals. During sampling, we use a 250-step Euler integrator, ensuring the same NFE as previous works such as REPA [43] and DiT [29]. To enhance sampling performance, we adopt cfg interval [19] and timestep shift similar to FLUX [20]. We benchmark our method against prior AR generation and latent diffusion approaches in Table 3.

We report four distinct sets of results, detailing the performance with and without cfg for both extended training (800 epochs) and rapid training (64 epochs). At 800 epochs, our model achieves state-of-the-art performance with an FID of 1.35. Furthermore, our model demonstrates exceptional performance in cfg-free generation, achieving an FID of 2.17, which surpasses the results of many methods that utilize cfg. Our approach also exhibits rapid convergence capabilities; at 64 epochs, it achieves an FID of 2.11, representing a speedup of over 21 times compared to the original

DiT. This further underscores the superiority of our method.

6. Ablations and Discussions

In this section, we perform ablation experiments on the design of VF loss to assess the impact of various foundation models and loss formulations. We then provide a deeper analysis of the underlying mechanism of VF loss, offering additional insights that might be helpful.

6.1. Generative Friendly VA-VAE

We demonstrate that the VA-VAE with a patch size of 1 exhibits superior generative performance compared to the SD-VAE with a patch size of 2. As shown in Table 1, replacing the SD-VAE [33] with the VA-VAE results in a reduction of the FID-50k from 7.13 to 4.29. This improvement can be attributed to two main reasons. Firstly, we observe that the DiT trained with a tokenizer using f16 and a patch size of 1 converges more readily than the DiT using f8 and a patch size of 2. Secondly, the vision foundation model is capable of enhancing its generative performance while maintaining its reconstruction fidelity.

6.2. Alations on Vision Foundation Models

We train our VA-VAE using three types of foundation models: self-supervised models [15, 28] with masked image modeling, the image-text contrastive learning model CLIP [32], and the Segment Anything model [18]. As in Section 5, we set w_{hyper} to 0.1, with $m_1 = 0.5$ and $m_2 = 0.25$. To accelerate convergence, we adjust the learning rate and global batch size to 1e-4 and 256, respectively. In contrast to previous settings, each tokenizer is trained on ImageNet 256×256 for 50 epochs. For each tokenizer, we train LightningDiT-B in the corresponding latent space for 160 epochs. Table 4 summarizes our findings, showing that all these vision foundation models enhance the generative performance of diffusion models. Among them, the self-supervised pre-trained model DINOv2 achieves superior generative results.

6.3. Ablations on Loss Formulations

We conduct ablation experiments on the loss functions proposed in Section 3. In these experiments, we use DINOv2 as the vision foundation model to train the *f16d32* tokenizers for 50 epochs, comparing the reconstruction and generative results with different settings. We individually remove the margin cosine similarity loss (mcos), margin distance matrix similarity loss (mdms), and the margin from the loss function. Due to the presence of adaptive weighting, when we use a single loss individually, we halve the hyper weight to ensure a fair comparison. For all three scenarios, we observe a certain degree of performance degradation, which validates the effectiveness of these components.

Model Type	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓
naive	0.26	28.59	0.089	0.80	22.62
DINOv2 [28]	0.28	27.96	0.096	0.79	**15.82**
MAE [15]	0.28	28.33	0.091	0.80	19.89
SAM [18]	0.26	28.31	0.091	0.80	19.80
CLIP [32]	0.33	28.39	0.091	0.80	18.93

Table 4. **Ablation on Foundation Models.** We evaluate the impact of different VF losses on generative performance. Our results show that DINOv2 achieves the highest generative performance.

Loss Type	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓
NaN	0.26	28.59	0.089	0.80	22.62
full	0.28	27.96	0.096	0.79	**15.82**
-mcos loss	0.27	28.52	0.090	0.80	21.87
-mdistmat loss	0.27	28.24	0.090	0.80	17.74
-margin	0.27	28.07	0.093	0.79	17.77

Table 5. **Ablation Study of VF Loss Formulations**: Comparison of different configurations on generative performance metrics using LightningDiT-B.

6.4. Discuss on VF loss with Latent Distribution

In discrete visual tokenizers, there is also a conflict between reconstruction and generation [42, 47]. A clear indicator of this conflict is codebook utilization. When the codebook is scaled up, reconstruction performance improves, but codebook utilization significantly decreases, resulting in uneven distribution in the discrete space.

We observe a similar phenomenon in continuous tokenizers. Specifically, we use t-SNE [40] to visualize the distribution of different latent spaces. Figure 6 shows that VF loss effectively improves the uniformity of the distribution. This observation is further supported by calculating the standard deviation and Gini coefficients of data point distribution using kernel density estimation (KDE) in Tabel 6. The uniformity metric of the tokenizer seems to be positively correlated with the generative gFID. As the uniformity metric improves, the generative performance of the corresponding tokenizer also increases.

7. Conclusion

This paper focuses on the optimization dilemma in latent diffusion systems. To address the problem, we propose VA-VAE, a VAE aligned with vision foundation models, and LightningDiT, an optimized DiT incorporating advanced design strategies. In VA-VAE, the VF loss function—comprising marginal cosine similarity and distance matrix losses—aligns the VAE's latent space with the vision model, resulting in a more uniform feature distribution and up to 2.8× faster convergence. With Light-

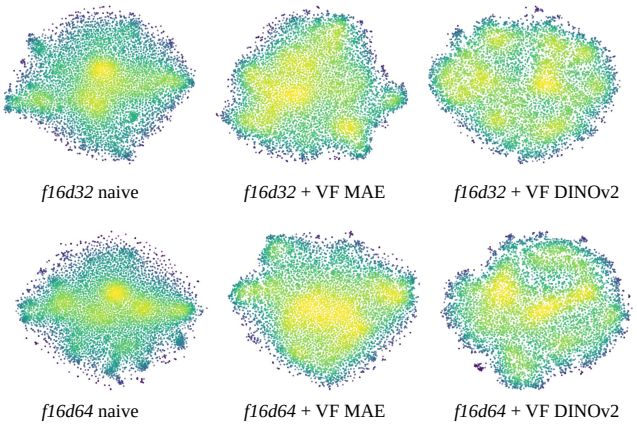

Figure 6. **Visualization of latent space with t-SNE.** VF loss makes the latent space distribution of high-dimensional tokenizers more uniform.

Tokenizer	VF Loss	density cv↓	gini coefficient↓	normalized entropy↑	gFID (DiT-B)↓
f16d32	*NaN*	0.263	0.145	0.995	22.62
	MAE	0.193	0.101	0.997	19.89
	DINOv2	**0.178**	**0.096**	**0.998**	**15.82**
f16d64	*NaN*	0.296	0.166	0.994	36.83
	MAE	0.256	0.143	0.995	**23.58**
	DINOv2	**0.251**	**0.141**	**0.996**	24.00

Table 6. **Relationship between uniformity and generative performance**: We evaluate the uniformity of feature distribution. Results indicate a possible positive correlation between the uniformity of feature distribution and generative performance.

ningDiT, we integrate advanced training techniques and architectural improvements to achieve faster DiT convergence. Combining the high-reconstruction capability of VA-VAE (rFID=0.28) with the rapid convergence of LightningDiT, our approach achieves a state-of-the-art FID of 1.35 on ImageNet 256. Besides, our method achieves 2.11 FID with only 64 epochs, demonstrating 21.8× speedup to original DiT. To the best of our knowledge, it is the first time that a latent diffusion system could achieve superior reconstruction and generation performance without additional training costs. We hope our work could help following research on latent diffusion systems.

References

[1] Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Zhen Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. Intriguing properties of quantization at scale. *Advances in Neural Information Processing Systems*, 36:34278–34294, 2023. 4

[2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 3, 7

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3, 4

[4] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 2, 3, 6

[5] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 1, 2

[6] 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, 4, 5

[7] 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, 2, 3, 4, 5, 6

[8] Fal. AuraFlow-v0.3. `https://huggingface.co/fal/AuraFlow-v0.3`. Accessed: 2024-10-28. 5

[9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 5

[10] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 3

[11] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23164–23173, 2023. 3, 5, 7

[12] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 7

[13] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pages 393–411. Springer, 2025. 2

[14] 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. 5

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 2, 5, 8

[16] Maciej Kilian, Varun Jampani, and Luke Zettlemoyer. Computational tradeoffs in image synthesis: Diffusion, masked-token, and next-token prediction. *arXiv preprint arXiv:2405.13218*, 2024. 2

[17] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2, 3

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 8

[19] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024. 5, 7

[20] Black Forest Labs. Frontier ai lab. https://blackforestlabs.ai/. Accessed: 2024-11-12. 2, 5, 7

[21] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024. 5, 6, 7

[22] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 2, 3

[23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4, 5

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[25] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[26] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024. 3, 5, 7

[27] OpenAI. Sora. https://openai.com/sora, 2024. 3, 4

[28] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3, 5, 8

[29] 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. 2, 3, 4, 6, 7

[30] 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. 2

[31] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 4

[32] 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. 3, 8

[33] 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, 2, 3, 4, 5, 6, 7, 8

[34] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 4, 5

[35] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4, 5

[36] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 6, 7

[37] PyTorch Team. Torch Compile Tutorial. https://pytorch.org/tutorials/intermediate/torch_compile_tutorial.html. Accessed: 2024-10-28. 5

[38] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 7

[39] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[41] Jingfeng Yao, Wang Cheng, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *arXiv preprint arXiv:2410.10356*, 2024. 4, 5, 7

[42] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh

Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 2, 3, 7, 8

[43] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 3, 5, 7

[44] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. 4, 5

[45] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 3, 7

[46] 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. 5

[47] Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*, 2024. 2, 3, 8