Beyond External Guidance: Unleashing the Semantic Richness Inside Diffusion Transformers for Improved Training

Lingchen Sun[1,2], Rongyuan Wu[1,2], Zhengqiang Zhang[1,2], Ruibin Li[1],
Yujing Sun[1,2], Shuaizheng Liu[1,2], Lei Zhang[1,2*]
[1]The Hong Kong Polytechnic University   [2]OPPO Research Institute

Abstract

*Recent works such as REPA have shown that guiding diffusion models with external semantic features (e.g., DINO) can significantly accelerate the training of diffusion transformers (DiTs). However, this requires the use of pretrained external networks, introducing additional dependencies and reducing flexibility. In this work, we argue that DiTs actually have the power to guide the training of themselves, and propose **Self-Transcendence**, a simple yet effective method that achieves fast convergence using internal feature supervision only. It is found that the slow convergence in DiT training primarily stems from the difficulty of representation learning in shallow layers. To address this, we initially train the DiT model by aligning its shallow features with the latent representations from the pretrained VAE for a short phase (e.g., 40 epochs), then apply classifier-free guidance to the intermediate features, enhancing their discriminative capability and semantic expressiveness. These enriched internal features, learned entirely within the model, are used as supervision signals to guide a new DiT training. Compared to existing self-contained methods, our approach brings a significant performance boost. It can even surpass REPA in terms of generation quality and convergence speed, but without the need for any external pretrained models. Our method is not only more flexible for different backbones but also has the potential to be adopted for a wider range of diffusion-based generative tasks. The source code of our method can be found at https://github.com/csslc/Self-Transcendence.*

1. Introduction

Diffusion models have emerged as a powerful framework for generative learning, achieving remarkable performance across a wide range of tasks, including image generation [10, 19], video synthesis [25, 30, 41, 48], and multi-modal

applications [7, 8, 23, 33]. Despite the great success, training diffusion transformers (DiTs) [26, 32] remains computationally intensive and suffers from slow convergence. Many methods [13, 16, 17, 21, 43, 45, 46, 50, 53, 55] have been developed to stabilize the DiT model training and accelerate the convergence process. Recent studies [43, 50] have highlighted the crucial role of semantically meaningful intermediate representations in both improving training efficiency and enhancing generative capability.

To enrich feature representations, several representation learning strategies have been proposed, including masked training [11, 12, 53, 54], contrastive learning [43], and representation alignment [21, 42, 50]. Among them, the pioneering work REPA [50] introduces an effective regularization strategy to align DiT features with external pretrained vision encoders such as DINO [31], significantly accelerating model training and improving image generation performance. However, such success comes at the cost of relying heavily on external networks, which introduces additional dependencies and reduces the flexibility of applications to different backbones [26, 32, 49].

To eliminate the reliance on external supervision, recent works [13, 16, 43] have explored self-contained alternatives. Dispersive Loss [43] introduces a plug-and-play regularizer that encourages feature dispersion without requiring pre-training or auxiliary data. SRA [16] and LayerSync [13] instead leverage the discriminative features in deeper layers during training to guide the learning of shallower layers. Specifically, SRA employs the EMA model as a teacher during training and performs layer-wise distillation across different noise levels. LayerSync introduces a lightweight synchronization mechanism to align semantically richer and weaker layers. Both of them eliminate the need for pre-trained external feature extractors. However, their performance still lags behind externally-guided approaches such as REPA [50].

We argue that this performance gap stems from the weaker discriminative information in internally generated guidance signals. As illustrated in the right side of Fig. 1, when synthesizing the image of a bird, REPA effectively
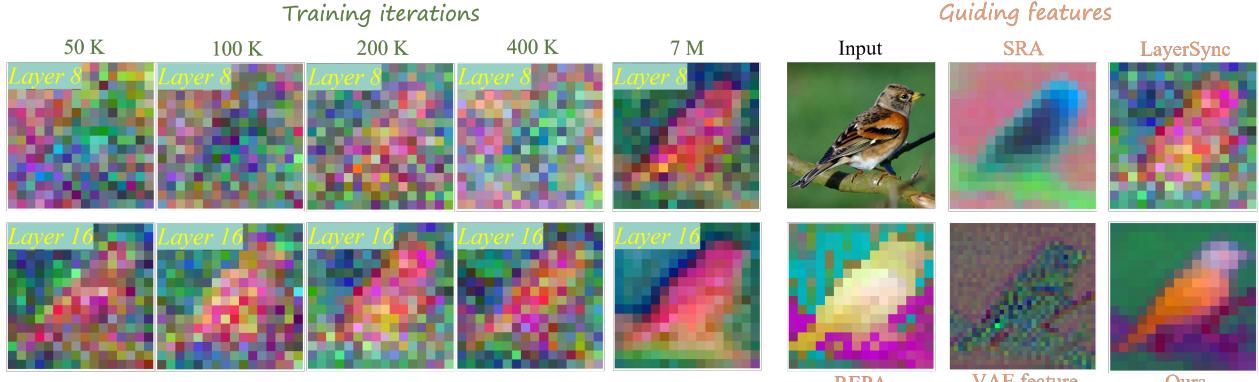
Figure 1. **Left**: PCA visualization [1] of latent features from both shallow (layer 8) and deeper (layer 16) blocks of SiT with $t = 0.6$ during training. Both layers progressively learn clean and discriminative representations, but the shallow layer learns such representations at a slower pace compared to the deeper one. **Right**: Comparison of guiding features from different methods. Our proposed approach produces clearer structural organization and more semantically richer features, as pre-trained DINO [31] used in REPA [50].

highlights semantically meaningful regions, leveraging features from the pre-trained DINO model. In contrast, the internal features used by SRA and LayerSync, derived from the model under training, lack a strong enough semantic representation, making them less effective in guiding shallow layers. This observation suggests that the semantic potential inherent in the DiT architecture has not been fully exploited. Therefore, a critical question arises: *Can internal features be used as effective semantic guidance signals to improve the training of DiT models?*

In this work, we attempt to answer this question and introduce **Self-Transcendence**, a simple yet effective self-guided training strategy achieving REPA-level performance without any external feature supervision. On the left side of Fig. 1, we visualize the latent features of a shallow (layer 8) and a deeper (layer 16) DiT block using PCA [1] at different training stages. Both layers gradually learn more discriminative patterns over time, but the shallow layer progresses very slowly. This indicates that the slow convergence of DiT is mainly due to the difficulty in learning clean and semantically rich features in shallow layers.

Based on the above observations, we propose a two-stage pipeline to explore more effective and semantically enriched internal features to guide the training of shallow blocks. (1) **VAE-based Alignment**: As shown on the right side of Fig. 1, self-contained methods like SRA and LayerSync use features from the model itself as guidance. However, these guiding features are updated along with training and tend to be noisy and unstable in early stages, lacking a clear semantic structure and the ability to effectively guide shallow layers. To address this, we directly use the clean latent features from the VAE as a stable and semantically meaningful guide, helping the model build structured internal representations in the early stage. However, while providing a good starting point, the semantic richness of VAE

features is limited. This motivates our second stage of (2) **Self-guided Representation Alignment**: After a warm-up phase, we extract intermediate features from deeper layer of the partially trained model and apply classifier-free guidance (CFG) to help the model generating stronger internal semantics, as shown in Fig. 1.

This two-stage strategy not only enjoys the benefits of the standard latent diffusion framework (*i.e.*, VAE feature), but is also entirely self-contained, achieving self-transcendence for training DiT models. While training such a guiding model incurs some overhead, it greatly accelerates the subsequent DiT training. Unlike the use of external vision encoders in prior work [42, 50], our approach does not require external data, and it is overall more efficient. Our method offers a new perspective, showing that the internal features from the DiT model itself can serve as effective guidance. Our contributions are summarized as follows:

- We introduce Self-Transcendence, a fully self-guided training framework to improve DiT model training without relying on external features.
- We propose a two-stage pipeline to obtain a semantically richer internal representation, which first aligns shallow-layer with VAE latent features, and then enhances intermediate features using CFG.
- We show that our method achieves on par performance with REPA in both convergence and generation quality, while having more flexibility for different backbones.

2. Related Work

Architectures of Diffusion Models. Early diffusion models typically adopt a U-Net backbone [34], consisting of convolution and attention layers [40]. Recently, Diffusion Transformers (DiTs) [32] replace U-Net with pure transformer blocks, following the design of Vision Transformers
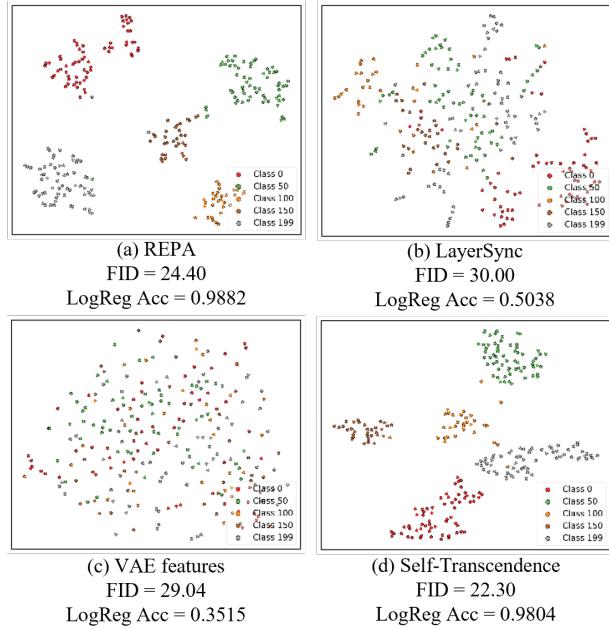
Figure 2. t-SNE visualizations of the guiding features extracted from (a) REPA [50], (b) LayerSync [13], (c) VAE features, and (d) our Self-Transcendence with $t = 0.4$ in the 200K iteration of SiT-XL/2. Different colors represent different classes. As REPA, our internal guiding features demonstrate superior class separability.

[9]. To further enhance DiT, Scalable Interpolant Transformers (SiT) [26] introduce a more flexible interpolant framework to generalize the diffusion process, systematically exploring the design choices of time discretization, prediction targets, interpolant types, and sampling strategies. LightningDiT [49] pushes DiT to its performance limits by incorporating a range of training and architectural optimizations, such as RMSNorm [51], SwiGLU [36], and RoPE [37], enabling faster convergence and more efficient inference. Other efforts investigate architectural improvements such as U-shaped transformer designs [38], dynamic computation adjustment [52], mixture-of-experts [2], linear attention mechanisms [47], and decoupled transformer design [44] to further boost the scalability and efficiency of diffusion models. Our proposed Self-Transcendence presents a new training strategy for the DiT family, guiding the model itself to converge faster.

Accelerating DiT Training. Recent studies [21, 50] highlight the importance of semantically meaningful representations for improving both training efficiency and generation quality. MaskDiT [53] accelerates DiT training by randomly masking 50% of input patches, encouraging efficient learning of informative features. MAETok [4] applies the masking strategy to tokenizer training, improving diffusion models by learning a semantically structured latent space through masked autoencoding. RCG [24] learns a generative model that generates semantic representations extracted by a self-supervised encoder, using them as conditions for image generation. REPA [50] introduces a simple representation alignment loss that aligns internal features with pretrained image embeddings [31], significantly boosting training speed and generation performance. Following works have begun to explore the use of pretrained vision encoders to provide richer external supervision. U-REPA [39] extends REPA to the U-Net architecture. VA-VAE [49] addresses the optimization bottleneck in latent diffusion by aligning the latent space of a tokenizer with pretrained vision foundation models, enabling faster convergence and better generation quality in high-dimensional settings.

In contrast, some recent works aim to accelerate training without pretrained vision encoders. Dispersive Loss [43] promotes diverse internal representations during diffusion training without external feature guidance. SRA [16] and LayerSync [13] guide shallow layers using semantically richer internal features. However, these methods still lag behind REPA in terms of performance. To bridge this gap, we propose Self-Transcendence, which leverages DiT model's own representations as a substitute for external features.

3. Method

3.1. Motivation and Framework Overview

With the widespread adoption of DiTs [26, 32, 49], how to accelerate their training process has become an increasingly important research topic [38, 44, 49, 50, 53]. Generally speaking, shallow blocks of DiT models are responsible for discriminative tasks, *i.e.*, separating clean latent states from noise in the given noisy input. On the other hand, deeper blocks focus on refining details based on the representations provided by the shallow layers [44]. However, as training progresses, a challenge emerges: learning discriminative features in shallow layers is significantly slower (see Fig. 1) due to the long gradient propagation path. This observation has motivated a line of research [13, 16, 21, 43, 44, 50] that explores how to train shallow blocks to better learn discriminative representations.

One promising approach is to introduce a guiding feature for supervising the shallow feature. REPA [50] initiates this research by introducing external features obtained from DINO [31] to guide shallow layers. DINO is a self-supervised vision encoder that learns powerful semantic representations without labels. As shown in Fig. 2(a), DINO embeddings form clear clusters, indicating strong semantic separability. These features effectively enhance the learning of shallow layers in the DiT models, thus significantly accelerating the DiT training. However, relying on external features like DINO introduces new dependencies: it requires extensive and time-consuming pre-training with external data, which may not always be feasible and desirable. Recent works have begun to explore whether diffusion

models can achieve self-acceleration. For example, Layer-sync and SRA [13, 16] use deeper layer features to guide the learning of shallower features. While these approaches are self-contained, their features lack stable structural guidance and semantic separability for training, leading to weaker performance, as shown in Fig. 2(b).

To better understand the requirements of guiding features, we experiment using VAE features. Although VAE lacks the strong discriminative power, its latent space is clean and structured. Surprisingly, using VAE features can accelerate the training to a level comparable to LayerSync. This suggests that a clean structure alone can still provide effective guidance, even without high discriminability. Of course, if the features also have strong discriminative power, the guidance will become more effective, as demonstrated by the superior results of REPA.

Combining these insights, we reveal that the most effective guiding features should meet two criteria: (1) *they should have a clean structure, in the sense that they can effectively help shallow blocks distinguish noise from signal*, and (2) *they should be semantically discriminative, making it easier for shallow layers to learn effective representations.* Several works [6, 20, 22, 27] have pointed out that diffusion models can be leveraged for various discriminative tasks, such as classification. This further motivates us to think whether we can find a more discriminative and semantically enriched feature representation inside the DiT model to guide the training of itself, achieving self-transcendence.

With these considerations, we propose a two-stage training framework, namely **Self-Transcendence**, as illustrated in Fig. 3. Firstly, we use clean VAE features as guidance to help the model distinguish useful information from noise in shallow layers. After a certain number of iterations, the model has learned more meaningful representations. We then freeze this model and use its representation as a fixed teacher, which avoids unstable guidance during training [13, 16, 43]. To enhance the semantic expression in the features, we build a self-guided representation that better aligns with the target conditions. Compared to REPA using external DINO features for supervision, our method replaces DINO features with features from a partially trained internal model, which demonstrates highly effective guiding performance. While this warm-up stage introduces the overhead of pretraining a guiding model, it significantly accelerates the subsequent model training and improves performance, making the extra cost worthwhile. In the next subsections, we introduce the two training stages of Self-Transcendence in detail.

3.2. VAE-based Alignment

Standard diffusion models are trained with a single denoising loss applied at the output, which often leads to slow learning in shallow layers. To improve this, we introduce

(a) VAE-based alignment
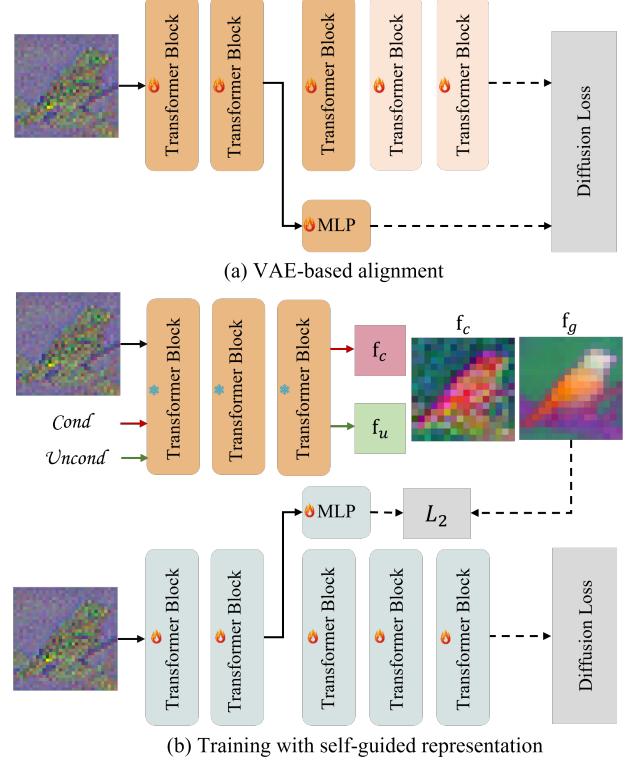
(b) Training with self-guided representation

Figure 3. The framework of our proposed Self-Transcendence. The spark icon indicates that the parameters of this layer are trainable, while the snowflake icon indicates that they are frozen.
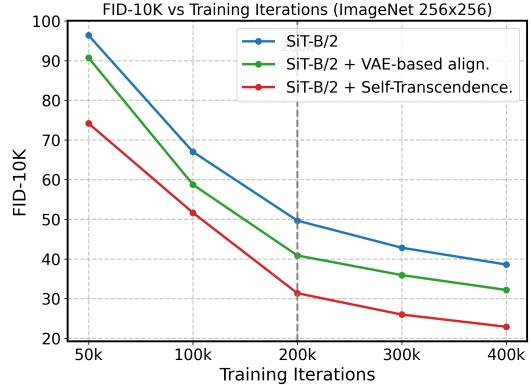

Figure 4. Comparison of FID-10K scores across training iterations on ImageNet (256×256). VAE-based alignment accelerates SiT training, while leveraging this model for self-transcendence leads to further improvements.

an auxiliary VAE-based alignment loss at intermediate layers, as shown in Fig. 3(a). Specifically, at the n-th layer, we extract the intermediate feature $\mathbf{f}_n$, pass it through a lightweight multilayer perceptron (MLP), and align it with the ground-truth VAE latent $\mathbf{z}$ using an L_2 loss, as shown in Eq. (1). The latent representation from the VAE provides a clean structural prior that helps the shallow layers distinguish meaningful signals from noisy inputs.

4

$$\mathcal{L}_{\text{VAE-align}} = \|\text{MLP}(\mathbf{f}_n) - \mathbf{z}\|_2^2. \tag{1}$$

This alignment can be regarded as an additional diffusion loss constrained only in shallow layers, facilitating them to better perceive structural information and speed up model convergence. As shown in Fig. 4, this simple alignment alone can already accelerate the learning of DiT, even obtaining better performance than the existing self-contained methods (please refer to Table 1). However, VAE features are generally less discriminative than semantic features extracted from external models like DINO. Therefore, while VAE-based alignment improves structural learning, it is not sufficient for semantic alignment. To address this, we propose a self-guided representation next.

3.3. Self-guided Representation

During the training of the diffusion model over time, the internal features gradually become more discriminative. Previous works have shown that features from deeper layers often capture stronger semantic information [13, 16, 50]. Therefore, we use deep-layer features as guiding signals to improve the training of the model, as shown in Fig. 3(b). However, as shown in Figs. 2(a) and (b), even with 200K training steps, the semantic richness of these features still lags behind that of external vision encoders like DINO.

To address this, we draw inspiration from Classifier-Free Guidance (CFG) [15], a technique widely used in conditional diffusion models. CFG improves the alignment between generated samples and the conditioning input without requiring an external classifier. During training, the model randomly removes the condition to learn both conditional and unconditional denoising. In the inference stage, it combines predictions from both modes using a guidance scale:

$$\epsilon_{\text{CFG}} = \epsilon_\theta(x_t, t, \phi) + \omega \cdot (\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \phi)), \tag{2}$$

where x_t is the sample state at timestep t, $\epsilon_\theta(x_t, t, c)$ and $\epsilon_\theta(x_t, t, \phi)$ denote diffusion predictions with and without the condition. Increasing w strengthens the influence of the conditioning signal, often generating images that better match the desired condition.

Building on this idea, we extend the CFG from the output space to the feature space. We extract both conditional and unconditional features from the same layer under the same input. Then the two features are combined using a guidance scale, as shown in Eq. (3):

$$\mathbf{f}_g = \mathbf{f}_u + \omega \cdot (\mathbf{f}_c - \mathbf{f}_u), \tag{3}$$

where $\mathbf{f}_c$ and $\mathbf{f}_u$ are the conditional and unconditional features, respectively. Eq. (3) encourages internal representations to align more closely with the desired semantics. As illustrated in Fig. 3(b) and Fig. 2(d), this feature-level guidance highlights the semantic region and significantly

improves the discriminative separability of deep features $\mathbf{f}_g$ compared to their original counterparts $\mathbf{f}_c$.

We then use the semantically enriched feature $\mathbf{f}_g$ to guide the shallower layers together with the standard diffusion loss, as shown in Fig. 3(b). Specifically, the intermediate features are passed through a lightweight MLP, and an L_2 loss is applied to align them. We combine the standard diffusion loss with our self-guided loss to optimize the diffusion model, i.e., $\mathcal{L} = \mathcal{L}_{diff} + \lambda_{guide} \times \mathcal{L}_{guide}$, where

$$\mathcal{L}_{guide} = \|\text{MLP}(\mathbf{f}_n) - \mathbf{f}_g\|_2^2. \tag{4}$$

Our approach explores the intrinsic discriminative ability of diffusion models by applying CFG on deep features. This is used to improve the overall discriminative power of the shallow layers in DiT, further accelerating DiT training, as illustrated in Fig. 4.

4. Experiments

4.1. Experimental Setup

Implementation Details. We conduct experiments using two baseline models: SiT [26] with a patch size of 2, and LightningDiT [49] with a patch size of 1. To ensure a fair comparison, we follow the default training and inference settings provided by each baseline. For the proposed Self-Transcendence method, several components should be determined: the guiding model, the guidance scale ω in the self-guided representation, loss weight λ_{guide}, as well as the guided and guiding layers. For all baselines, we use the model trained with the VAE-based alignment method at 40 epochs as the guiding model. Note that the guiding model shares the same architecture as the baseline model. We set the guidance scale to $\omega = 30.0$ and $\omega = 10.0$ for the SiT and LighteningDiT backbones, respectively. To choose the guided and guiding layers, we consider the total number of Transformer blocks n in each model. We select the guided layer as $n/2$ and the guiding layer as $2n/3$. We provide the ablation studies about these parameters in Sec. 4.3.

We apply an early stop strategy during training. The self-guided representation loss is only used during the early iterations (20 epochs for base models, 10 epochs for larger models). After these iterations, we remove the self-guided loss and only optimize the diffusion loss. This is because self-transcendence aims to improve the training of shallow layers, which lack semantic structures. Over-training the shallow layers can make the training of the deeper layers unstable and harm the modeling of the joint data distribution. Similar phenomena have been observed in REPA [45]. Further discussions are provided in Sec. 4.4.

Evaluation Metrics. To evaluate the quality of generated samples, we employ standard evaluation metrics [26, 50]. We use the Fréchet Inception Distance (FID) [14] as the primary metric, as we also report sFID [29], inception score

[35], precision, and recall [18]. All metrics are calculated on 50,000 samples unless otherwise stated.

Compared Methods. We compare with different acceleration methods: REPA [50], Disperse Loss [43], SRA [16], and LayerSync [13]. Various latent diffusion models are also compared. (1) U-Net backbone: LDM [34]. (2) Hybrid Transformer and U-Net backbone: U-ViT-H/2 [3] and MDTv2-XL/2 [12]. (3) Transformer backbone: MaskDiT [53], SD-DiT [54], DiT-XL/2 [32], and SiT-XL/2 [26].

4.2. Main Results

Comparison with Existing Acceleration Methods. Table 1 shows the comparison between our proposed VAE-based alignment and Self-Transcendence methods with other acceleration approaches. The following observations can be made. **(1)** Firstly, VAE-based alignment alone achieves competitive results with existing self-contained methods such as Disperse Loss and LayerSync. This is because semantic structure is mainly learned during the early training stages. However, Disperse Loss and LayerSync fail to provide strong and stable guidance in this phase. In contrast, our alignment features are derived from a pretrained VAE, and they are fixed during training, making it more stable and easier to guide the learning process. **(2)** Secondly, our proposed Self-Transcendence method significantly outperforms all other self-contained techniques. On SiT-XL/2, our method achieves 7.51 FID with only 80 epochs, outperforming the LayerSync trained for 200 epochs (8.80 FID). In addition, Self-Transcendence achieves results comparable to or even better than REPA, which uses external DINO features. **(3)** Thirdly, our approach also brings substantial improvements to LightningDiT, showing that it can be generalized to different backbones (SiT and DiT) and different VAE latent spaces (SD-VAE[34] and VAVAE). Notably, on LightningDiT-XL/1, Self-Transcendence achieves an FID of 3.55 with only 64 epochs of training.

Finally, although our method requires a pretrained guiding model using VAE-based alignment, this overhead is minimal and worthwhile. As shown in SiT-B/2 and SiT-XL/2, it outperforms the longer-trained baselines by a large margin. This demonstrates that a small cost in warmup training can lead to significant acceleration and performance gains. Moreover, compared to the widely used guiding model (DINO), the training for our guiding model is much easier and more efficient without using external data.

Scalability. Scalability, which refers to the ability to maintain or improve performance as model size or data scale increases, is an important property for a training paradigm. We evaluate the proposed Self-Transcendence across different model sizes. As shown in Table 1, our method consistently accelerates training at all scales. Notably, the performance gain becomes larger as the model size increases. For example, on SiT/B-2, the FID improves from 36.14 (base-

Table 1. Comparisons with different acceleration methods based on the vanilla SiTs and LightningDiTs on ImageNet 256×256. CFG is not used. $\downarrow$ denotes that lower values are better.

Model	#Params	Epochs	FID$\downarrow$
SiT-B/2	130M	120	31.45
SiT-B/2	130M	80	36.14
+ REPA	130M	80	24.40
+ Disperse Loss	130M	80	32.45
+ LayerSync	130M	80	30.00
+ VAE-based alignment (Ours)	130M	80	29.04
+ Self-Transcendence (Ours)	130M	80	**20.49**
SiT-L/2	458M	80	21.41
+ REPA	458M	80	9.70
+ Disperse Loss	458M	80	16.68
+ LayerSync	458M	80	14.83
+ VAE-based alignment (Ours)	458M	80	14.61
+ Self-Transcendence (Ours)	458M	80	**8.74**
SiT-XL/2	675M	800	8.30
SiT-XL/2	675M	120	14.74
SiT-XL/2	675M	80	17.63
+ REPA	675M	80	7.90
+ Disperse Loss	675M	200	10.64
+ LayerSync	675M	200	8.80
+ VAE-based alignment (Ours)	675M	80	12.25
+ Self-Transcendence (Ours)	675M	80	**7.51**
LightningDiT-B/1 (w. VAVAE)	130M	64	15.94
+ VAE-based alignment (Ours)	130M	64	15.87
+ Self-Transcendence (Ours)	130M	64	**14.03**
LightningDiT-XL/1 (w. VAVAE)	675M	64	5.30
+ REPA	675M	64	4.09
+ VAE-based alignment (Ours)	675M	64	5.07
+ Self-Transcendence (Ours)	675M	64	**3.55**

line) to 20.49 with a 43.3% relative reduction. On SiT/XL-2, it improves from 17.63 to 7.51, with a 57.4% reduction. This clearly demonstrates the scalability of our method. In addition, our guiding model shares the same backbone with the generation model, which may benefit from larger generation models. Therefore, our approach may have more potential for even better acceleration on higher-capacity diffusion models, which we leave for future study.

Visualization of Training Process. We compare the vanilla SiT model, REPA-enhanced model, and our trained model across 100K to 400K iterations on two ImageNet classes, as illustrated in Fig. 5. Both our model and REPA show faster convergence than the vanilla SiT model, and our method tends to obtain better structural generation. For example, in the shoe class, our method generates more realistic shapes and consistent textures earlier in training. By 400K iterations, our model yields sharper contours and finer details, indicating improved sample quality and stability during training. This can be owed to the fact that the guiding feature and the trained model in our method share a similar architecture and learning process, so that the shallow layers

Figure 5. Visual comparison of generated samples from SiT-XL/2 models at different training iterations. For all models, we apply the same seed, noise, and sampling strategy with a CFG scale of 4.0.

Table 2. Comparisons across diffusion backbones and acceleration methods on ImageNet 256×256 using CFG. ↓ and ↑ indicate whether lower or higher values are better, respectively.

Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
U-Net						
LDM-4	200	3.60	-	247.7	0.87	0.48
Transformer + U-Net hybrid						
U-ViT-H/2	240	2.29	5.68	263.9	0.82	0.57
MDTv2-XL/2	1080	1.58	4.52	**314.7**	0.79	0.65
Transformer						
MaskDiT	1600	2.28	5.67	276.6	0.80	0.61
SD-DiT	480	3.23	-	-	-	-
DiT-XL/2	1400	2.27	4.60	278.2	0.83	0.57
SiT-XL/2	1400	2.05	4.50	270.3	0.82	0.59
LightningDiT-XL/1	800	1.35	4.15	295.3	0.79	0.65
Training Acceleration						
SiT-XL/2						
+ REPA	800	1.42	4.70	305.7	0.80	0.65
+ Disperse Loss	≥1200	1.97	-	-	-	-
+ SRA	800	1.58	4.65	311.4	0.80	0.63
+ LayerSync	800	1.89	-	265.3	0.81	0.60
+ Ours	**400**	1.44	4.85	311.3	**0.79**	**0.66**
LightningDiT-XL/1						
+ Ours	**400**	**1.25**	**4.11**	303.9	0.78	**0.66**

are easier to learn the semantic and structural information.

Comparison of Diffusion Models using CFG. We quantitatively compare Self-Transcendence against recent latent diffusion models using two backbones, SiT-XL/2 and LightningDiT-XL. Leveraging the semantically rich latent space of VAVAE [49], our method outperforms all the compared diffusion models using only 400 epochs in the LightningDiT-XL backbone. For the SiT-XL/2 backbone, our method achieves similar performance to REPA with

CFG in most metrics at just 400 epochs, showing significant performance improvement compared over the other self-contained methods (SRA and LayerSync).

Table 3. Comparisons across diffusion backbones and acceleration methods on ImageNet 512×512 using CFG. ↓ and ↑ indicate whether lower or higher values are better, respectively.

Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Pixel diffusion						
VDM++	–	2.65	–	278.1	–	–
Latent diffusion, Transformer						
MaskDiT	800	2.50	5.10	256.3	0.83	0.56
DiT-XL/2	600	3.04	5.02	240.8	**0.84**	0.54
SiT-XL/2	600	2.62	4.18	252.2	**0.84**	0.57
+ REPA	200	2.08	4.19	274.6	0.83	0.58
+ Ours	100	2.00	**4.11**	265.0	0.83	0.58
+ Ours	200	**1.76**	4.16	**286.6**	0.82	**0.62**

To further evaluate the robustness and scalability of our method under higher-resolution scenarios, we conduct experiments on ImageNet at a resolution of 512×512. This setting poses greater challenges for both semantic alignment and visual fidelity due to the increased spatial complexity and richer details. We follow the same training pipeline as in the 256×256 experiments, except for the input resolution. Specifically, original images are encoded into $64 \times 64 \times 4$ latent representations using the VAE encoder from Stable Diffusion [34]. As shown in Table 3, the baseline SiT-XL/2 model [26] exhibits limited performance at this resolution, with a FID of 2.62 and IS of 252.2 after 600 training epochs. Incorporating REPA [50] significantly improves the generation performance, reducing FID to 2.08 and increasing IS to 274.6, with 200 epochs. Our

Self-Transcendence framework shows stronger generation acceleration than REPA, achieving an FID of 2.00, IS of 265.0 in only 100 epochs.

Visual Examples of Generated Images. We provide qualitative results of SiT-XL/2 on ImageNet at resolutions of 256 and 512 using the proposed Self-Transcendence method in Fig. 6 and Fig. 7, respectively, showcasing its ability to generate realistic structures and textures across diverse semantic categories at different resolutions. For animals, such as dog, panda, and bird, our method captures fine-grained details such as fur texture, feather patterns, and facial features with high fidelity. In natural scenes, such as the flower and stone, our trained models generate complex lighting, depth, and organic shapes with coherence and realism. For manmade and structured objects such as doll, wooden barrel, and pinwheel, the generated images exhibit accurate geometry and sharp edges.

Table 4. Ablation studies. All SiT-B-2 models are trained with 80 epochs. FID is calculated on 10,000 samples.

Methods	VAE-based align.	Self-guided rep.	FID↓	IS↑
SiT-B/2	✗	✗	38.60	41.95
V1	✓	✗	32.20	52.38
V2	✗	✓	25.21	63.83
Ours	✓	✓	**22.91**	**70.37**

4.3. Ablation Study

Effectiveness of Each Component. To evaluate the effectiveness of each component in Self-Transcendence, we conduct ablation studies by removing its two key components individually: VAE-based alignment and self-guided representation. The results are reported in Table 4. In variant V1, the model is trained only with the VAE-based alignment loss and the standard diffusion loss. Since VAE features lack rich semantics, removing the self-guided stage leads to slower convergence and suboptimal performance. In V2, we use the model trained solely with diffusion loss as the guiding model in the self-guided stage. Without the initial alignment by VAE features, the guiding model struggles to learn meaningful information within 40 epochs. Overall, the full model equipped with both components achieves the best performance, confirming the complementary roles of VAE-based alignment and self-guided representation in providing guidance and accelerating DiT training.

Hyperparameters. We conduct ablation studies on the selection of guiding model, loss wight λ_{guide}, guidance scale (see Eq. (3)), and guiding and guided layers. The results are shown in Table 5. The top row shows our default setting (layer 6 guided by layer 8 with guidance scale 30.0).

(1) Guiding and guided layers. We first investigate how the selection of guiding and guided layers affects performance. As shown in Table 5, we denote the configuration as '$m \rightarrow n$', meaning that the mth layer of the guid-

Table 5. Ablation studies on the guidance scale, guiding and guided layers, the selection of the guiding model, and the loss weight λ_{guide}. '$m \rightarrow n$' means the nth layer of DiT model is guided by the mth layer of guiding model. All the SiT/B-2 models are trained with 80 epochs. FID is calculated on 10,000 samples.

Block Layers	Guidance Scale	Guiding model, Training iters	λ_{guide}	FID↓	IS↑
$8 \rightarrow 6$	30.0	$M_{align.}$, 200K	0.5	22.91	70.37
$8 \rightarrow 4$	30.0	$M_{align.}$, 200K	0.5	23.43	70.25
$8 \rightarrow 8$	30.0	$M_{align.}$, 200K	0.5	24.29	66.88
$6 \rightarrow 6$	30.0	$M_{align.}$, 200K	0.5	24.37	67.89
$10 \rightarrow 6$	30.0	$M_{align.}$, 200K	0.5	24.05	67.94
$8 \rightarrow 6$	1.0	$M_{align.}$, 200K	0.5	29.30	55.68
$8 \rightarrow 6$	15.0	$M_{align.}$, 200K	0.5	24.26	68.36
$8 \rightarrow 6$	45.0	$M_{align.}$, 200K	0.5	23.01	70.69
$8 \rightarrow 6$	60.0	$M_{align.}$, 200K	0.5	23.57	68.20
$8 \rightarrow 6$	30.0	$M_{align.}$, 50K	0.5	28.32	57.22
$8 \rightarrow 6$	30.0	$M_{align.}$, 100K	0.5	25.20	64.05
$8 \rightarrow 6$	30.0	$M_{align.}$, 300K	0.5	22.91	71.32
$8 \rightarrow 6$	30.0	M_{ori}, 200K	0.5	25.21	63.83
$8 \rightarrow 6$	30.0	M_{repa}, 200K	0.5	23.40	70.81
$8 \rightarrow 6$	30.0	$M_{align.}$, 200K	0.1	24.28	67.50
$8 \rightarrow 6$	30.0	$M_{align.}$, 200K	0.3	23.08	71.41
$8 \rightarrow 6$	30.0	$M_{align.}$, 200K	0.7	23.27	70.12
$8 \rightarrow 6$	30.0	$M_{align.}$, 200K	1.0	23.24	68.97

ing model is used to supervise the nth layer of the current model. We observe that guiding shallow layers (*e.g.*, 8→6 and 8→4) consistently leads to better performance than guiding deeper layers like 8→8. One possible explanation is that deeper layers are closer to model output, thus overconstraining these layers may interfere with the model's ability to adapt to the data distribution. Guiding the same layer depth (6→6) results in noticeably worse performance, indicating that certain level of abstraction gap between the guiding and guided features is necessary. In addition, using very deep guiding layers (10→6) slightly under-performs 8→6, implying that very deep features are too semantically distant from shallow layers (as indicated in REPA [50]), which cannot provide effective guidance to them. Overall, these results highlight the importance of choosing guiding layers, which should be semantically strong and appropriately aligned with the target layers.

(2) Guidance scale. We then study the impact of the guidance scale used in the self-guided representation stage. To evaluate sensitivity, we vary the scale from 1.0 to 60.0. As shown in the middle part of the table, reducing the scale to 1.0 significantly degrades performance, indicating that weak guidance is insufficient for effective semantic transfer. Increasing the guidance scale to a stronger value of 45.0 slightly improves IS but does not outperform the default setting. A radical choice of guidance scale (60.0) results in degraded performance in both FID and IS. This may be because overly strong guidance introduces training instability or causes the model to overfit to intermediate features,
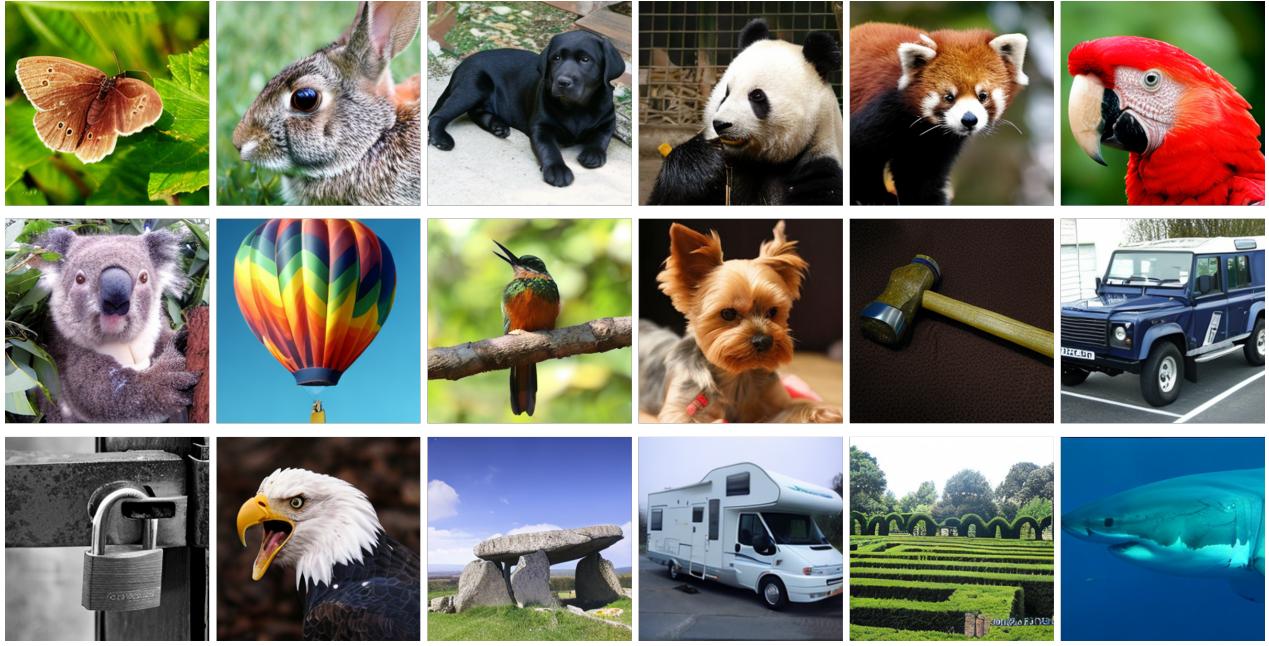
Figure 6. Examples of generated images on ImageNet 256×256 of our proposed Self-Transcendence method. We use classifier-free guidance with 4.0 scale.

Figure 7. Examples of generated images on ImageNet 512×512 of our proposed Self-Transcendence method. We use classifier-free guidance with 4.0 scale.

limiting final performance [5, 28]. We find that a moderate guidance scale (*i.e.*, 30.0) provides a good balance between stability and semantic enhancement.

(3) Guiding model. Thirdly, we evaluate how the guiding models affect performance. We denote the models trained with our VAE-based alignment, the original diffusion loss, and REPA as M_{align}, M_{ori}, and M_{repa}, respectively. First,

we vary the training iterations of the same guiding model (M_{align}) from 50K to 300K. As shown in Table 5, using a better trained guiding model leads to better FID scores. For example, with 50K training steps, the FID is 28.32, while at 200K steps, it improves to 22.91. However, further increasing training to 300K does not bring consistent gains (FID = 22.91). This implies that around 200K steps,
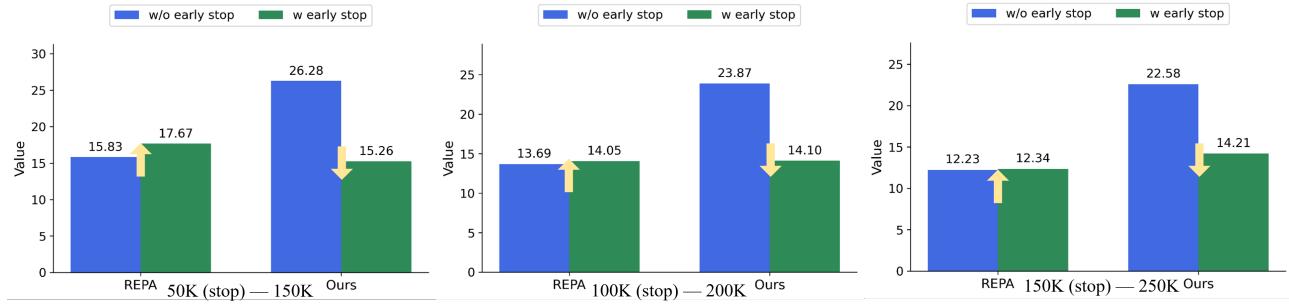
Figure 8. FID-10K scores with and without early stopping at various training stages of REPA and our Self-Transcendence method. For example, '50K (stop) — 150K' means that the alignment loss is not optimized after 50K iterations, and only the diffusion loss is optimized to 150K iterations. Both REPA and our method are more sensitive to early stopping in the earlier stages (*e.g.*, at 50K iterations). However, our Self-Transcendence method benefits from early stopping, achieving better FID scores, while REPA's performance degrades when the early stop strategy is applied.
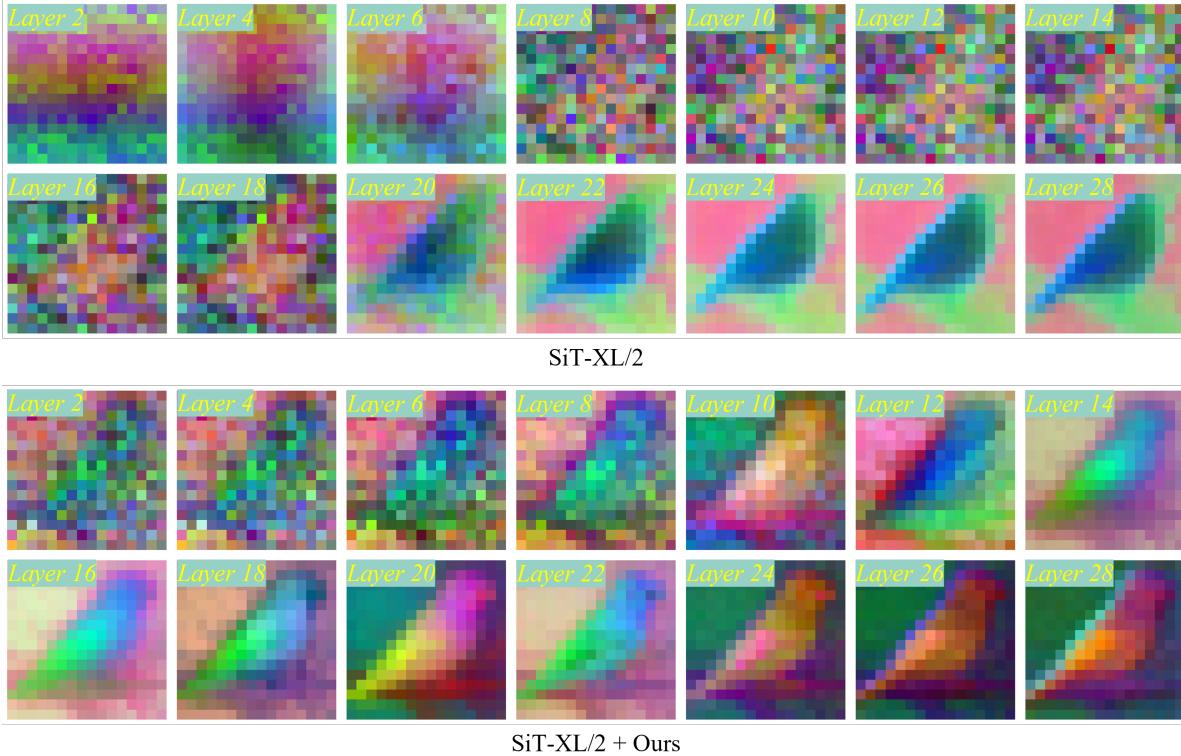

Figure 9. PCA feature visualization from different layers of SiT-XL/2 and SiT-XL/2+Self-Transcendence with 400K iterations and $t = 0.7$.

the intermediate features of the guiding model become semantically rich enough to offer useful supervision. With further training, its internal representations may shift from the current model. This is similar to the representation drift problem observed in knowledge distillation [5, 28], where a too-strong teacher may misguide the student. Second, we compare different types of guiding models. Using a model trained with standard diffusion loss (M_{ori}) results in worse performance, showing that the lack of VAE-based alignment reduces guiding quality. Meanwhile, using a model trained with REPA (M_{repa}) with a stronger perfor-

mance also achieves worse performance than ours, suggesting that the structural and semantic priors in our guiding model $M_{align.}$ provide more effective guidance.

(4) Loss weight. Finally, to investigate the impact of different loss components on the performance of Self-Transcendence, we conduct an ablation study on the loss weight. During training, we apply two loss functions: the diffusion loss $\mathcal{L}_{diff}$ and the self-guided loss $\mathcal{L}_{guide}$. The weight of the diffusion loss λ_{diff} is fixed to 1.0, while the weight of the self-guided loss λ_{guide} is varied from 0.1 to 1.0. The results are summarized in Table 5. We observe
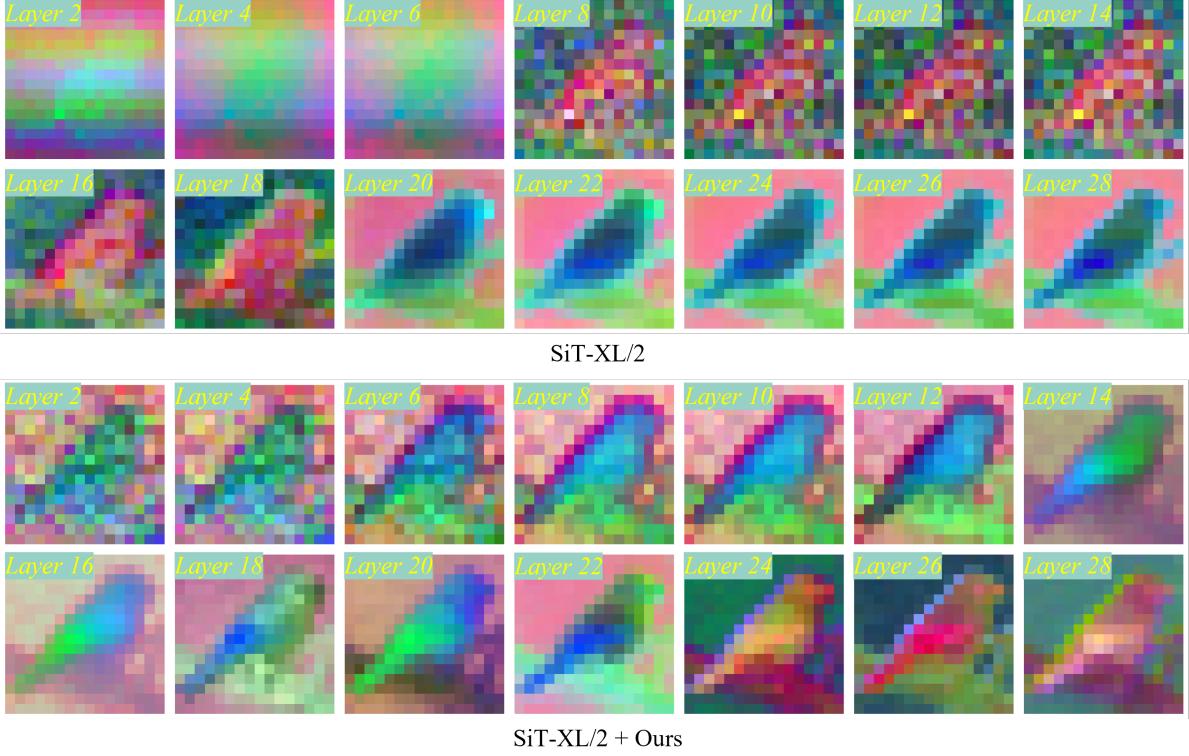
10

Figure 10. PCA feature visualization from different layers of SiT-XL/2 and SiT-XL/2+Self-Transcendence with 400K iterations and $t = 0.5$.

that setting the self-guided loss weight to 0.5 achieves the best overall performance, reaching a FID of 23.91 and IS of 70.37, outperforming other configurations across most metrics. When λ_{guide} is too small (*e.g.*, 0.1), the guidance effect is insufficient, leading to degraded generation quality (FID = 24.28). Conversely, increasing λ_{guide} (*e.g.*, to 1.0) results in performance deterioration (FID = 23.24), likely due to over-regularization. These results confirm the importance of balancing the two loss terms. A moderate guidance weight (*e.g.*, 0.5) provides optimal control without overwhelming the primary diffusion objective. Therefore, in our final model, we choose $\lambda_{guide} = 0.5$ as the default setting.

4.4. Early Stop Strategy

We investigate the effect of early stopping on our method during training. Specifically, we stop optimizing the self-guided loss after a certain number of iterations (*e.g.*, 50K) and continue training using only the diffusion loss for 100K more iterations (*e.g.*, 150K). This strategy is indicated by *50K (stop) — 150K*. All experiments are performed using the SiT-XL/2 backbone, and 10,000 samples are used for evaluation. As shown in Fig. 8, both REPA [50] and our Self-Transcendence method are more sensitive to early stopping at earlier stages. However, our method benefits from early stopping, achieving lower FID-10K scores, while REPA's performance degrades when the alignment loss is stopped early. One possible reason is that over-training the shallow layers may destabilize the training of deeper layers and hinder the modeling of joint data distribution. In contrast, our method can take advantage of reduced optimization, leading to a lower total computational cost.

4.5. Training Time Comparison

We then compare the training time of Self-Transcendence, REPA [50], and the vanilla model [26]. All the compared models are trained with 256 batch size. Using SiT-B/2 as an example, we train for 200 epochs (1,000K iterations) on 8 A800 GPUs. The vanilla model trains at 9.28 iters/s, taking approximately 29.93 hours. With REPA, the speed drops to 8.49 iters/s, resulting in a total training time of 32.72 hours. For our Self-Transcendence method, the first 50K iterations run at 6.69 iters/s and the remaining 950K at 9.28 iters/s, leading to a total training time of about 30.52 hours. Although slightly slower in the early phase, our method achieves a comparable overall training time to the vanilla model and is faster than REPA. Moreover, REPA relies on a pretrained DINOv2 model [31], which incurs significant additional cost. In contrast, training our guiding model takes about only 6.39 hours (200K iterations at 8.70 iters/s),

which is substantially more efficient than DINOv2. Therefore, Self-Transcendence offers a more resource-efficient alternative while maintaining strong performance.

5. Feature Map Visualization

We provide PCA visualizations [1] of feature maps to compare the evolution of representation across layers. As shown in Fig. 9 and Fig. 10, Self-Transcendence significantly enhances feature organization in different layers, showing more compact and structured representations. In contrast, the vanilla model exhibits more scattered and less coherent patterns, indicating weaker discriminative features.

6. Conclusion and Limitation

In this work, we proposed **Self-Transcendence**, a simple yet effective self-guided training framework to improve the training of diffusion transformers (DiTs). Unlike previous approaches that relied on external pretrained models for semantic supervision, our method was entirely self-contained and it leveraged the model's own internal features to guide its training. We designed a two-stage pipeline, where we first aligned shallow-layer features with VAE latents to provide stable early supervision and then applied classifier-free guidance to enhance the semantic expressiveness of intermediate features. The obtained features were used to guide a new DiT training. Extensive experiments demonstrated that our method achieved comparable or even superior performance to externally guided methods such as REPA, while offering greater flexibility for various DiT backbones. Our findings highlighted the untapped potential of internal representations in DiT models and provided a new direction for self-supervised acceleration.

Limitations. While our method eliminates the need for external models, it introduces an additional training stage to bootstrap internal semantics, which adds a small amount of overhead in the early phase. Moreover, the quality of internal guidance is still upper-bounded by the model's own capacity. Finally, as in previous works [13, 43, 44, 49], our approach is evaluated on class-to-image benchmarks; its applicability to other generative modalities (*e.g.*, text to image generation, text to video generation, and 3D generation) deserves to be explored in future work.

References

[1] Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. 2, 12

[2] Anonymous. Dense2moe: Unifying pruning and upcycling for efficient large language models. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. under review. 3

[3] Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion mod-

els. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. 6

[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 *International Conference on Learning Representations, year=2025,*. 3

[5] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4793–4801, 2019. 9, 10

[6] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 4

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[8] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[11] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer, 2023. 1

[12] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer, 2024. 1, 6

[13] Yasaman Haghighi, Bastien van Delft, Mariam Hassan, and Alexandre Alahi. Layersync: Self-aligning intermediate layers, 2025. 1, 3, 4, 5, 6, 12

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[16] Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is

needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 1, 3, 4, 5, 6

[17] Felix Krause, Timy Phan, Ming Gui, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Tread: Token routing for efficient architecture-agnostic diffusion training. *arXiv preprint arXiv:2501.04765*, 2025. 1

[18] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 6

[19] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. Official inference repo for FLUX.1 models. 1

[20] Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[21] 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. 2025. 1, 3

[22] Alexander Cong Li, Mihir Prabhudesai, Shivam Duggal, Ellis Langham Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. 4

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. 1

[24] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2024. 3

[25] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. 1

[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, 2024. 1, 3, 5, 6, 7, 11

[27] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. *Advances in Neural Information Processing Systems*, 37:55141–55177, 2024. 4

[28] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020. 9, 10

[29] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 5

[30] OpenAI. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024. 1

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 1, 2, 3, 11

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 1, 2, 3, 6

[33] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 1

[34] 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 (CVPR)*, pages 10684–10695, 2022. 2, 6, 7

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques

for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[36] Noam Shazeer. Glu variants improve transformer, 2020. 3

[37] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. 3

[38] Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped diffusion transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[39] Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits, 2025. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. 1

[42] Chenyu Wang, Cai Zhou, Sharut Gupta, Zongyu Lin, Stefanie Jegelka, Stephen Bates, and Tommi Jaakkola. Learning diffusion models with flexible representation guidance. *arXiv preprint arXiv:2507.08980*, 2025. 1, 2

[43] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization, 2025. 1, 3, 4, 6, 12

[44] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 3, 12

[45] Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint arXiv:2505.16792*, 2025. 1, 5

[46] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think, 2025. 1

[47] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024. 3

[48] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[49] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3, 5, 7, 12

[50] 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. In *International Conference on Learning Representations*, 2025. 1, 2, 3, 5, 6, 7, 8, 11

[51] Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada, 2019. 3

[52] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. *arXiv preprint arXiv:2410.03456*, 2024. 3

[53] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *Transactions on Machine Learning Research*, 2024. 1, 3, 6

[54] Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer, 2024. 1, 6

[55] Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer, 2024. 1