SD-DiT: Unleashing the Power of Self-supervised Discrimination in Diffusion Transformer*

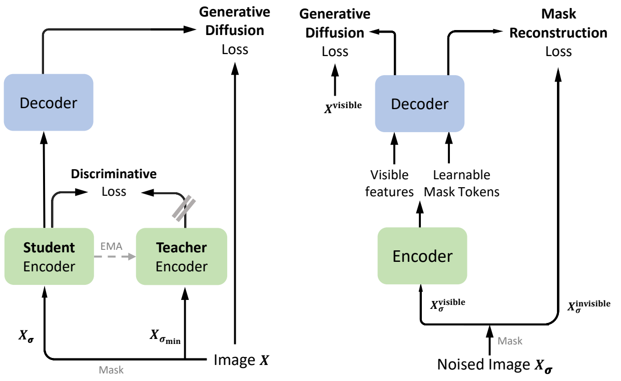
Rui Zhu[1], Yingwei Pan[2], Yehao Li[2], Ting Yao[2], Zhenglong Sun[1], Tao Mei[2], Chang Wen Chen[3]
[1] The Chinese University of HongKong, Shenzhen [2] HiDream.ai Inc. [3] The Hong Kong Polytechnic University
ruizhu@link.cuhk.edu.cn, {pandy, liyehao, tiyao}@hidream.ai, sunzhenglong@cuhk.edu.cn
tmei@hidream.ai, changwen.chen@polyu.edu.hk

Abstract

Diffusion Transformer (DiT) has emerged as the new trend of generative diffusion models on image generation. In view of extremely slow convergence in typical DiT, recent breakthroughs have been driven by mask strategy that significantly improves the training efficiency of DiT with additional intra-image contextual learning. Despite this progress, mask strategy still suffers from two inherent limitations: (a) training-inference discrepancy and (b) fuzzy relations between mask reconstruction & generative diffusion process, resulting in sub-optimal training of DiT. In this work, we address these limitations by novelly unleashing the self-supervised discrimination knowledge to boost DiT training. Technically, we frame our DiT in a teacher-student manner. The teacher-student discriminative pairs are built on the diffusion noises along the same Probability Flow Ordinary Differential Equation (PF-ODE). Instead of applying mask reconstruction loss over both DiT encoder and decoder, we decouple DiT encoder and decoder to separately tackle discriminative and generative objectives. In particular, by encoding discriminative pairs with student and teacher DiT encoders, a new discriminative loss is designed to encourage the inter-image alignment in the self-supervised embedding space. After that, student samples are fed into student DiT decoder to perform the typical generative diffusion task. Extensive experiments are conducted on ImageNet dataset, and our method achieves a competitive balance between training cost and generative capacity.

1. Introduction

Recent computer vision field has witnessed the rise of diffusion models [29, 43, 62, 67] in powerful and scalable generative architectures for image generation. Such practical generative model pushes the limits of a series of CV applications, including text-to-image synthesis [2, 53, 54, 57], video generation [7, 23, 30, 80], and 3D generation [15, 76].



(a) SD-DiT (b) MaskDiT

Figure 1. Conceptual comparison between (a) our SD-DiT and (b) MaskDiT. MaskDiT integrates generative diffusion process with mask reconstruction auxiliary task, and the whole DiT encoder plus decoder are jointly optimized for the two tasks. In contrast, our SD-DiT frames mask modeling on the basis of discrimination knowledge distilling in a self-supervised manner, pursuing the inter-image alignment in the joint embedding space of teacher and student encoder. DiT encoder and decoder are decoupled to separately tackle discriminative and generative diffusion objectives.

A recent pioneering practice is the Diffusion Transformer (DiT) [49], which inherits the impressive scaling properties of Transformers [72] and significantly improves the capacity & scalability of diffusion models. Unfortunately, similar to Vision Transformers [18], the training of DiT usually suffers from slow convergence and heavy computation burden issues. The recent works [21, 81] then turn their focus on investigating the way to accelerate the training convergence of DiT. Many consider combining the Transformer-based diffusion process with additional mask reconstruction objective via the popular mask strategy [10, 16, 25]. In particular, MDT [21] simultaneously encodes both the complete and masked image input, in order to enhance the intra-image contextual learning among the associated patches. MaskDiT [81] integrates generative diffusion process with mask reconstruction auxiliary task to optimize the whole DiT encoder and decoder (see Fig. 1b).

*This work was performed at HiDream.ai.

Although significantly improved training efficiency is attained, these DiT architectures with mask strategy still struggle with extremely high-fidelity image synthesis and suffer from several inherent limitations. (1) **Training-inference discrepancy**: Mask strategy inevitably introduces learnable mask tokens for triggering mask reconstruction during DiT training, but no artificial mask token is involved for generative diffusion process at inference. This training-inference discrepancy severely limits the generative capacity of learned DiT. Note that to alleviate such discrepancy, MDT introduces additional dual-path interaction between complete and masked inputs during training, while sacrificing much higher computational and memory cost. (2) **Fuzzy relations between mask reconstruction & generative diffusion process**: Most mask-based DiT structures process both the visible and learnable mask tokens via the same DiT decoder to jointly enable mask reconstruction and generative diffusion process, leaving the inherent different peculiarity of each objective not fully exploited. It is noteworthy that such mask modeling can be regarded as intra-image reconstruction derived from the same data distribution (*e.g.*, from $p_{\sigma \odot \text{mask}}$ to p_σ for noised data in MaskDiT). Instead, the generative diffusion process aims to model the translations between the real data distribution p_{data} and a different noised data distribution p_σ. This issue is also observed in MaskDiT, where mask reconstruction objective will gradually overwhelm generative objective at the late training stage. Accordingly, the joint training of the two distinct objectives with fuzzy relations results in sub-optimal training of DiT when applied to generative task.

To address these limitations, our work paves a new way to frame mask modeling of DiT training on the basis of discrimination knowledge distilling in a self-supervised fashion. We propose a novel Diffusion Transformer model with Self-supervised Discrimination, namely SD-DiT, that pursues highly-efficient learning of DiT with higher generative capacity. Technically, SD-DiT shapes the discrimination knowledge distilling in a teacher-student scheme. As shown in Fig. 1a, the input discriminative pairs of teacher and student DiT encoders are derived from different diffusion noises (*i.e.*, p_{σ_1} and p_{σ_2} along the same Probability Flow Ordinary Differential Equation (PF-ODE) of EDM [34]). More importantly, different from typical mask strategy that triggers mask reconstruction objective over both DiT encoder and decoder, SD-DiT decouples DiT encoder and decoder to separately perform discrimination knowledge distilling and generative diffusion process. Our launching point is to fully exploit the mutual but also fuzzy relations between self-supervised discrimination distillation and generative diffusion process through such decoupled DiT design. Eventually, we devise a new discriminative loss to enforce the inter-image alignment of encoded visible tokens between teacher and student DiT encoders in the joint embedding space. Next, SD-DiT only feeds student samples into student DiT decoder for performing the conventional generative diffusion objective. Note that here our discriminative loss can be interpreted as inter-image translation between teacher sample (approximately real data distribution p_{data}) and student sample (noised data distribution p_σ), which better aligns with generative diffusion objective than conventional intra-image mask reconstruction objective. As such, the joint optimization of discriminative and generative diffusion objectives strengthens DiT training both effectively and efficiently.

In the meantime, the student branch (student DiT encoder plus decoder) in our decoupled DiT design completely retains the same regular noise in EDM and modules as in the generative modeling at inference. The additional teacher DiT encoder is simply updated as the Exponential Moving Average (EMA) of student DiT encoder in a light-weight fashion, without incurring a heavy computational burden for self-supervised discrimination. In this way, our SD-DiT not only preserves the training efficiency of mask modeling, but also elegantly circumvents the training-inference discrepancy issue.

The main contribution of this work is the proposal of Diffusion Transformer structure that fully unleashes the power of self-supervised discrimination to facilitate DiT training. This also leads to the elegant view of how a training-efficient DiT architecture should be designed for fully exploiting the mutual but also fuzzy relations between mask modeling and generative diffusion process, and how to bridge the training-inference discrepancy tailored to generative task. Through extensive experiments on ImageNet-256×256, we demonstrate that our SD-DiT consistently seeks a better training speed-performance trade-off when compared to state-of-the-art DiT models.

2. Related Work

Diffusion Models. Denoising diffusion probabilistic models (DDPMs) [29] greatly accelerate the development of generative models, especially the tasks of text conditioned image synthesis [2, 46, 53, 54, 57], image editing [6, 12, 26, 42, 44, 48] and personalized image generation [20, 56]. As a score-based model [65, 66], DDPMs introduce a forward process to gradually add Gaussian noise to the data according to Stochastic Differential Equation [67], and the iterative denoising procedures are employed to generate high-quality samples. To tackle such a time-consuming iterative nature of DDPM, fast sampling strategies [28, 34, 41, 58, 63] and training diffusion in the latent space [54, 73] are proposed. Besides, several innovations for improving the network architecture of diffusion models are attained to handle various challenging generation tasks. Convolutional UNet [55] is the de-facto configuration from recent diffusion models [29] and

ADM [17] further boosts the generation quality of UNet with scalable model size, including the adaptive group normalization [75], the attention blocks [60, 70] and the residual blocks from BigGAN [5].

Diffusion Transformers. Transformers [72] provide a new paradigm to connect various domains across language [16], vision [4, 25, 37, 39, 78, 79, 82], and multimodalities [38, 52], with remarkable scaling properties in terms of model size [33] and pre-training efficiency [25]. Recently, some Transformer-based diffusion models [3, 32, 49, 77] are proposed to exploit the advantages of Transformer architecture in diffusion models. For example, GenViT [77] first presents that Vision Transformer (ViT) [18] has the potential for image generation. Based on ViT with long skip connections, U-ViT [3] is specifically designed for the diffusion model which is characterized by integrating the time, the specific condition, and the noisy image patches as tokens. DiT [49] systematically studies the scaling behaviors of Transformers under the Latent Diffusion Models (LDMs) [54] framework, and achieves better generation quality than the U-Net counterparts with a scaling-up high-capacity backbone. In this work, we take the conventional DiT blocks as backbone network and the generative diffusion task is implemented as LDMs.

Self-supervised Learning with Diffusion Models. With the dominant status of Transformers in vision and language, the mask strategy from self-supervised learning [4, 16, 25] has greatly propelled the development of generative models. Following the paradigm of bidirectional generative modeling [51], MaskGiT [10] and MUSE [11] aim at predicting randomly masked visual tokens which were first tokenized from images by a discrete VQ-VAE [19, 71]. Iterative decoding is further utilized to rapidly generate an image. Moreover, MAGE [36] employs such masked token modeling to unify representation pre-training and image generation. On the other hand, diffusion models built upon Transformers [18, 49] could be well integrated with the mask image modeling [74]. For example, inspired by MAE [25], MDT [21] and Mask-DiT [81] take advantage of the asymmetrical encoder-decoder of MAE and add the learning objective loss of reconstructing masked tokens (without discrete tokenizers) to the original generative diffusion loss. Such combination with mask modeling remarkably improves the training efficiency and the contextual reasoning ability of the Diffusion Transformer (*i.e.*, DiT). It is noteworthy that mask modeling is built upon the intra-view reconstruction while the typical self-supervised methods with discriminative joint embedding pretraining [8, 9, 13, 14, 22, 24] focus on the inter-view alignment (invariance). Different from existing mask strategy with intra-image contextual learning, our SD-DiT paves a new way to endow mask modeling in DiT with self-supervised discrimination ability via inter-image alignment.

3. Approach

In this paper, we devise a Diffusion Transformer with Self-supervised Discrimination (SD-DiT) to frame mask modeling in efficient DiT training as self-supervised discrimination knowledge distilling. This section starts with a brief review of the preliminaries of diffusion models. Then, the overall decoupled architecture for discriminative and generative objectives is elaborated. After that, two different kinds of objectives for generative diffusion process and mask modeling, *i.e.*, generative loss and discriminative loss, are introduced. Finally, the overall objective of SD-DiT at the training stage is provided.

3.1. Preliminaries

Diffusion models introduce a forward process to progressively add Gaussian noise to the data distribution $p_{\text{data}}(\boldsymbol{x})$ by a Stochastic Differential Equation (SDE) [67] over time:

$$d\boldsymbol{x}_t = \boldsymbol{\mu}(\boldsymbol{x}, t)dt + g(t)d\boldsymbol{\omega}_t, \tag{1}$$

where $\boldsymbol{\mu}$ and g are the drift and diffusion coefficients, and $\boldsymbol{\omega}$ is the standard Brownian motion. With the time flowing from 0 to T, we denote the marginal distribution of $\boldsymbol{x}_t$ as $p_t(\boldsymbol{x})$. Based on such an SDE, Song *et al.* [67] define the probability flow ordinary differential equation (PF-ODE) in the reverse-time sample generation process:

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{\mu}(\boldsymbol{x}, t) - \frac{1}{2}g(t)^2\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)]\mathrm{d}t. \tag{2}$$

Recent EDM [34] proposes to add Gaussian noise with mean zero and standard deviation σ into the data distribution. Specifically, EDM utilizes $p_\sigma(\boldsymbol{x})$ instead of $p_t(\boldsymbol{x})$ and configures $\boldsymbol{\mu}(\boldsymbol{x}, t) := \mathbf{0}$ and $g(t) := \sqrt{2t}$ in Eq. (2). In this case, the resulting perturbed distribution is given by $p_\sigma(\boldsymbol{x}) = p_{\text{data}}(\boldsymbol{x}) * \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where $*$ denotes the convolution operation. In other words, the real data $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x})$ can be directly diffused as:

$$\boldsymbol{x}_\sigma = \boldsymbol{x}_0 + \boldsymbol{n},\ \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}). \tag{3}$$

And the corresponding PF-ODE in EDM is presented as:

$$d\boldsymbol{x} = -\sigma\nabla_{\boldsymbol{x}} \log p_\sigma(\boldsymbol{x})d\sigma, \quad \sigma \in [\sigma_{\text{min}}, \sigma_{\text{max}}], \tag{4}$$

where $\nabla_{\boldsymbol{x}} \log p_\sigma(\boldsymbol{x})$ is the score function [67]. As such, diffusion models are basically regarded as score-based generative models [63, 65, 67]. To avoid numerical instability in ODE solving, σ_{min} is a small positive value and thus $p_{\sigma_{\text{min}}}(\boldsymbol{x}) \approx p_{\text{data}}(\boldsymbol{x})$, while σ_{max} is large enough so that $p_{\sigma_{\text{max}}}(\boldsymbol{x})$ is close to a tractable Gaussian distribution. The training objective of EDM is to minimize the expected L_2 denoising loss for $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x})$ separately for each σ, by parameterizing a denoiser network as D_θ:

$$\mathbb{E}_{\boldsymbol{x}_0 \sim p_{\text{data}}} \mathbb{E}_{\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})} \|D_\theta(\boldsymbol{x}_0 + \boldsymbol{n}, \sigma) - \boldsymbol{x}_0\|_2^2. \tag{5}$$

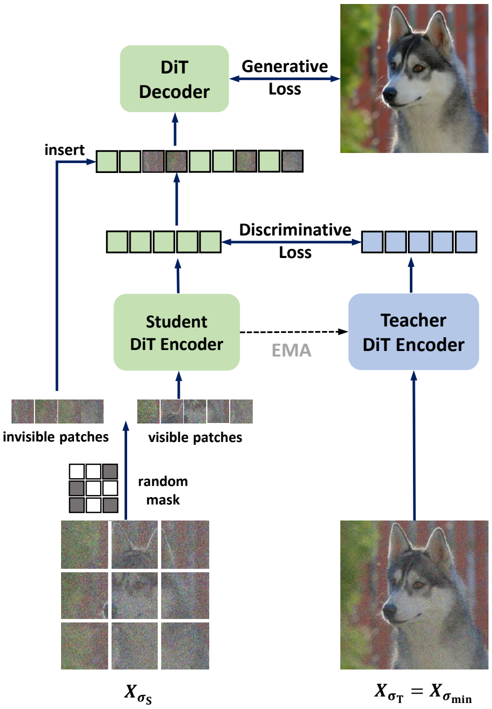


Figure 2. The overview of our SD-DiT. During training, the student view is diffused with regular noise as in EDM formulation, while the teacher view is derived from fixed minimum noise of the consistency function that is close to real data distribution. SD-DiT feeds the discriminative pair into teacher and student DiT encoders to perform self-supervised discriminative process within the joint embedding space. Meanwhile, only the student DiT encoder and DiT decoder undertake the generative diffusion process. At inference, all patches are fed into student branch for sampling.

The estimated score function is thus measured as:

$$\nabla_{\boldsymbol{x}} \log p_\sigma(\boldsymbol{x}) = (D_\theta(\boldsymbol{x}_\sigma, \sigma) - \boldsymbol{x}_\sigma)/\sigma^2. \qquad (6)$$

Based on the formulation of EDM, Consistency Models [64, 68] propose to learn a *consistency function* whose outputs of arbitrary pairs on the PF-ODE trajectory (Eq. (4)) are consistent with $\boldsymbol{x}_{\sigma_{\min}} \sim p_{\sigma_{\min}}(\boldsymbol{x}) \approx p_{\text{data}}(\boldsymbol{x})$. Formally, the *consistency function* is defined as:

$$\boldsymbol{f} : (\boldsymbol{x}_\sigma, \sigma) \mapsto \boldsymbol{x}_{\sigma_{\min}}, \qquad (7)$$

and reflects an important property of *self-consistency*:

$$\boldsymbol{f}(\boldsymbol{x}_\sigma, \sigma) = \boldsymbol{f}(\boldsymbol{x}_{\sigma'}, \sigma'), \quad \sigma, \sigma' \in [\sigma_{\min}, \sigma_{\max}]. \qquad (8)$$

The diffusion noising schedule in our SD-DiT follows the basic formulation of EDM. And the discrimination objective in our SD-DiT is framed on the basis of the theory of the *consistency function* (Eq. (7)).

3.2. Overall Architecture

The motivation of our SD-DiT is to exploit self-supervised discrimination to facilitate the efficient training of Diffusion Transformer. Fig. 2 illustrates the overall architecture of SD-DiT, which triggers mask modeling as discrimination knowledge distilling in a teacher-student scheme. **Decoupled Encoder-Decoder Structure.** Technically, our SD-DiT consists of teacher/student DiT encoders and one DiT decoder, and the core generative objective is framed on the basis of latent space as LDM [54]. The additional discriminative objective is shaped as inter-image alignment among teacher and student DiT encoders in self-supervised joint-embedding space [1, 9]. Considering the fuzzy relations between mask modeling and generative diffusion process, here we leverage a decoupled encoder-decoder structure to perform the joint training of generative and discriminative objectives, rather than optimizing the whole encoder-decoder with mask reconstruction objective as in existing methods [81]. Specifically, SD-DiT feeds the discriminative pairs into teacher and student DiT encoders to conduct discrimination knowledge distilling. After that, only student samples are fed into student DiT decoder to perform generative diffusion process. In this decoupled design, the discriminative objective only updates DiT encoder by empowering it with inter-image discriminative capacity. Meanwhile, DiT decoder is solely optimized with generative objective by retaining the same regular noise to nicely mimic the generative diffusion process at inference.

Discriminative Pairs. In an effort to trigger discriminative objective, we construct the input discriminative pairs based on the EDM formulation (Eq. (3)). Since the student branch (including student DiT encoder and decoder) will perform both the generative and discriminative objectives, here the student view should be diffused regularly within a large range, similar to MaskDiT [81]: $\boldsymbol{x}_{\sigma_S} = \boldsymbol{x}_0 + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma_S^2 \mathbf{I})$, $\sigma_S \in [\sigma_{\min}, \sigma_{\max}]$. For the teacher view, we take inspiration from the InfoMin principle [69] in self-supervised learning, and choose the fixed minimum noise of the *consistency function* [64, 68] to construct input samples: $\boldsymbol{x}_{\sigma_T} = \boldsymbol{x}_0 + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\min}^2 \mathbf{I})$. As such, the noised distribution of teacher view can be the closest one to the original data distribution ($\boldsymbol{x}_{\sigma_T} \sim p_{\sigma_{\min}}(\boldsymbol{x}) \approx p_{\text{data}}(\boldsymbol{x})$) and far away from the noised student view. Note that we empirically evaluate various teacher noise across $[\sigma_{\min}, \sigma_{\max}]$ in Sec. 4.4, and attain the similar observations as in InfoMin principle [69]: The noised teacher view too close to the noised student view could be harmful to self-supervised discriminative learning. Accordingly, we use the fixed minimum noise for teacher view in practice.

3.3. Generative Objective

Inspired by the training efficiency and location contextual awareness [21, 81] brought by mask strategy, we follow the typical mask modeling techniques (*e.g.*, [25]) to frame the generative objective via asymmetric encoder-decoder structure along the student branch.

Mask Strategy. The image will be divided into n non-overlapping patches through the patch embedding layer of DiT. Let $\mathcal{M}$ denote the binary random mask with the same size of non-overlapping patches. It is worth noting that MAE and MaskDiT additionally leverage the mask $\mathcal{M}$ to learn additional mask tokens in mask reconstruction auxiliary task. Instead, our SD-DiT solely utilizes the mask $\mathcal{M}$ to separate the noised student view into visible patches ($\boldsymbol{v}_{\sigma_{\rm S}} = \boldsymbol{x}_{\sigma_{\rm S}} \odot (1-\mathcal{M})$) and invisible patches ($\bar{\boldsymbol{v}}_{\sigma_{\rm S}} = \boldsymbol{x}_{\sigma_{\rm S}} \odot \mathcal{M}$), where $\odot$ indicates element-wise multiplication on patches.

Student Branch. Given the visible and invisible patches via mask strategy, the student branch applies the typical asymmetric encoder-decoder architecture [25, 81] to improve the training efficiency. The student DiT encoder can be built with various DiT-Small/Base/XL backbones, while the lightweight student DiT decoder consists of a fixed number of blocks (*i.e.*, 8 DiT blocks, similar to the configurations of MAE [25].). The student DiT encoder $\mathcal{S}_\theta$ only operates over the visible patch and obtains the visible tokens $\mathcal{S}_\theta(\boldsymbol{v}_{\sigma_{\rm S}})$. Then the student decoder $\mathcal{G}_\theta$ is fed with the complete token set $\mathcal{H}$. Such an asymmetric paradigm with a high mask ratio proposed by MAE [25] greatly reduces the training cost because the main computation burden is carried on the large-scale encoder.

Generative Loss. Recall that in existing mask modeling techniques (*e.g.*, MAE and MaskDiT), the input token set $\mathcal{H}$ of decoder commonly augments the visible tokens $\mathcal{S}_\theta(\boldsymbol{v}_{\sigma_{\rm S}})$ with learnable mask tokens, according to the positions of the mask $\mathcal{M}$. The mask reconstruction auxiliary task is included to recover the learnable mask tokens from the invisible patches $\bar{\boldsymbol{v}}$. It is noteworthy that such mask reconstruction objective can benefit the representation learning, but leaves the inherent different peculiarity of mask modeling and generative objectives under-exploited. MaskDiT also points out the fuzzy relations between these two objectives, where mask reconstruction loss will gradually overwhelm the generative objective at the late training stage.

To alleviate this limitation, we discard the mask reconstruction loss and optimize the DiT decoder with only generative loss. Formally, for the complete token set $\mathcal{H}$, we remove the learnable mask tokens and directly insert the invisible patches $\bar{\boldsymbol{v}}$ onto the visible tokens $\mathcal{S}_\theta(\boldsymbol{v}_{\sigma_{\rm S}})$, according to the positions of mask $\mathcal{M}$. Next, the generative loss operates over the compete tokens, which is measured in the form of EDM (Eq. (5)):

$$\mathcal{L}_{\rm G} = \mathbb{E}_{\boldsymbol{x}_0 \sim p_{\rm data}} \mathbb{E}_{\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\rm S}^2 \mathbf{I})} \| D_\theta(\boldsymbol{x}_0 + \boldsymbol{n}, \sigma_{\rm S}, \mathcal{M}) - \boldsymbol{x}_0 \|_2^2, \quad (9)$$

where D_θ denotes student branch including the student DiT encoder $\mathcal{S}_\theta$ and DiT decoder $\mathcal{G}_\theta$.

3.4. Discriminative Objective

Unlike typical mask modeling with mask reconstruction loss, our SD-DiT paves a new way to frame mask modeling of DiT training on the basis of discrimination knowledge distilling in a self-supervised manner. Inspired by self-distilling loss in ViT-based self-supervised methods (*i.e.*, DINO [9] and iBOT [82]), we design discriminative loss to enforce the inter-image alignment of encoded visible tokens between teacher and student DiT encoders.

Specifically, the teacher sample $\boldsymbol{x}_{\sigma_{\rm T}}$ is fed into teacher DiT encoder $\mathcal{T}_{\theta'}$, yielding the output tokens $\mathcal{T}_{\theta'}(\boldsymbol{x}_{\sigma_{\rm T}})$. Next, SD-DiT performs discriminative loss over the visible tokens between teacher $\boldsymbol{e}_{\rm T} = \mathcal{T}_{\theta'}(\boldsymbol{x}_{\sigma_{\rm T}})$ and student $\boldsymbol{e}_{\rm S} = \mathcal{S}_\theta(\boldsymbol{v}_{\sigma_{\rm S}})$ in the joint encoding space. A three-layer projection head j_θ operates on $\boldsymbol{e}_{\rm S}$ and $\boldsymbol{e}_{\rm T}$ and outputs the softmax probability distribution over K dimensions. By denoting the distribution on each student and teacher token as $P_{{\rm S}_i}$ and $P_{{\rm T}_i}$ ($i \in (1-\mathcal{M})$ indicates the index of visible tokens.), the softmax probability distribution of student is measured as:

$$P_{{\rm S}_i} = \frac{\exp(j_\theta(\boldsymbol{e}_{{\rm S}_i})/\tau_{\rm S})[k]}{\sum_{k=1}^{K} \exp(j_\theta(\boldsymbol{e}_{{\rm S}_i})/\tau_{\rm S})[k]}, \quad (10)$$

where the student temperature $\tau_{\rm S}$ controls the sharpness of the softmax distribution. A similar formulation also holds for teacher: $P_{{\rm T}_i}$ with teacher temperature $\tau_{\rm T}$. For each visible token i, the discrimination loss targets aligning the distribution between teacher and student by minimizing the cross-entropy loss:

$$\mathcal{L}_{\rm D}(i) = -\sum_k P_{{\rm T}_i} \log(P_{{\rm S}_i}). \quad (11)$$

The final discrimination loss is calculated over all visible patch tokens and the [CLS] token:

$$\mathcal{L}_{\rm D} = \frac{1}{(1-\mathcal{M})} \sum_{i \in (1-\mathcal{M})} \mathcal{L}_{\rm D}(i) + \mathcal{L}_{\rm D}(\texttt{[CLS]}). \quad (12)$$

Besides, we adopt the centering technique in DINO to avoid feature collapse, where the batch mean statistic is used to whiten the features before softmax during each training iteration. For simplicity, here we leave the details of centering and the complete pseudo-codes to supplementary materials.

In summary, the overall training loss is the combination of discrimination loss and generative loss: $\mathcal{L}_{\rm D} + \mathcal{L}_{\rm G}$. The parameters of student branch (student DiT encoder and decoder) are optimized by this overall loss. And teacher DiT encoder (parameterized as $\mathcal{T}_{\theta'}$) is updated as the exponential moving average (EMA) of student DiT encoder: $\mathcal{T}_{\theta'} = \beta \mathcal{T}_{\theta'} + (1-\beta)\mathcal{S}_\theta$. Here β is a momentum coefficient. During training, the teacher is updated by EMA without SGD back-propagation, thereby only requiring extremely lightweight computational cost. At inference, the teacher is completely removed and no burden is introduced.

4. Experiments

4.1. Implementation Details

In this section, we provide the settings of model architecture, training setup, and evaluation details. We list the detailed configurations in supplementary material.

Model Architecture. The basic Transformer blocks in our backbone network fully adopt the DiT [49] block which fuses conditional time and class embedding with adaptive layer normalization [50]. We follow the paradigm of LDM [54] and DiT to perform diffusion generation in the latent space of the frozen pre-trained VAE model [54], which downsamples a $256 \times 256 \times 3$ image into a $32 \times 32 \times 4$ latent variable. Inspired by [25, 81], we adopt the asymmetric encoder-decoder for generative diffusion process. The student DiT encoder $\mathcal{S}_\theta$ employs DiT-Small/Base/XL-2 (patch size: 2) and the small-scale DiT decoder $\mathcal{G}_\theta$ contains 8 DiT blocks, similar to the configurations of MAE. For the discriminative objective, we mainly follow the settings of iBOT [82] and DINO [9]. The teacher DiT encoder $\mathcal{T}_{\theta'}$ is the EMA of student encoder, and the momentum coefficient increases from 0.996 to 0.999 at the end of training. The three-layer projection head j_θ outputs the `[CLS]` and patch tokens with $K = 8,192$ dimension for softmax probability distribution in discriminative loss Eq. (11).

Training Setup. Following previous Transformer-based diffusion models [21, 49, 81], we conduct all the experiments on ImageNet-1K with 256×256 resolution and a batch size of 256. We adopt the most common settings of DiT, *e.g.*, AdamW [40] optimizer with a constant $1e-4$ learning rate and no weight decay. Without specified stating, the mask ratio is set to 0.2 on the student view, and no mask is applied on the teacher view. No data augmentation is employed for both student and teacher inputs since our model will learn the discrimination among various noised views. Notice that the mixed precision might lead to nanloss during training, so we only apply mixed precision for evaluation on small scale backbone (DiT-S) and transfer to full precision for large scale backbone (DiT-B and DiT-XL). All experiments are conducted on 8× 80GB-A100 GPUs.

Evaluations. To evaluate both the diversity and quality of our generative model, we utilize the most commonly adopted Fréchet Inception Distance (FID) [27] as evaluation metric. For fair comparison with previous works [21, 49, 81], we report FID-50K from ADM's TensorFlow evaluation suite [17] with the reference batch. We report the FID scores of the class-conditional sampling. Besides, we provide more supporting metrics including Inception Score (IS) [59], sFID [45] and Precision/Recall [35].

4.2. Training Speed *vs.* Performance

Here we evaluate our SD-DiT with regard to both training speed and generative performance. Fig. 3 shows the training speed (*i.e.*, training steps per second) and FID-50K score of SD-DiT in comparison to state-of-the-art DiT models (DiT [49], MDT [21], and MaskDiT [81]) on 8 × A100 GPUs. For fair comparison, the backbone network of each run is built on the same scale of DiT-S/2, same batch size (256) and training iterations (400k). For SD-DiT and MaskDiT, we follow MDT and implement them with the same Float32 precision. For a comprehensive analysis, we also label each run with the number of input patches. As shown in Fig. 3, our SD-DiT (FID: 48.39; speed: 9.2 steps/sec or 0.11 sec/step) obtains better generative performance with faster training speed than MDT (FID: 53.46; speed: 2.4 steps/sec or 0.42 sec/step) and DiT (FID: 68.40, 5.03 steps/sec or 0.20 sec/step). This is due to that MDT simultaneously forwards and backwards both the complete (100%) and visible patches (70%), and DiT operates over the complete (100%) patches, thereby resulting in slower training speed. In contrast, our SD-DiT and MaskDiT only forward and backward partial patches (80%/50%), leading to faster training speed. Furthermore, unlike MaskDiT that optimizes the whole encoder-decoder with mask reconstruction objective, our SD-DiT adopts a decoupled encoder-decoder structure to better exploit the mutual but also fuzzy relations between generative and discriminative objectives, leading to the best FID-50K score. The results basically demonstrate the effectiveness of our SD-DiT which seeks a competitive training speed-performance trade-off.

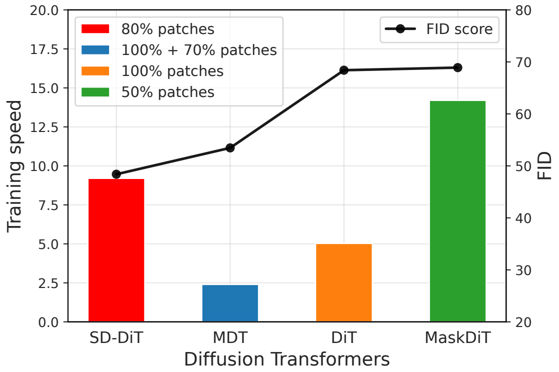


Figure 3. Training speed (training steps per second) *vs.* generative performance (FID-50K score) for our SD-DiT, MDT, DiT, and MaskDiT on 8 × A100 GPUs. We also label each run with the number of input patches.

4.3. Performance Comparison

Comparison among Backbones in Different Scales. Tab. 1 provides comprehensive comparisons between our SD-DiT and several DiT-based state-of-the-arts under three different model sizes (DiT-S/B/XL). Notice that Mask-DiT only conducts experiments on DiT-XL backbone so we do not report its results on DiT-S and DiT-B backbones. The batch size of all models is set as 256 for fair comparison. Specifically, under the same small-scale backbone (DiT-S), our SD-DiT-S (48.39) exhibits better performance than DiT-S (68.40) and MDT-S (53.46) by a large margin. This significant performance improvement of FID score is consistently observed when transferring to the larger scale backbones (DiT-B, DiT-XL). The results clearly validate the ad-

Method	Training Steps(k)	FID-50K↓
DiT-S/2 [49]	400	68.40
MDT-S/2 [21]	400	53.46
SD-DiT-S/2	400	**48.39**
DiT-B/2 [49]	400	43.47
MDT-B/2 [21]	400	34.33
SD-DiT-B/2	400	**28.62**
DiT-XL/2 [49]	7000	9.62
MaskDiT-XL/2 [81]	1300	12.15
MDT-XL/2 [21]	1300	9.60
SD-DiT-XL/2	1100	9.66
SD-DiT-XL/2	1300	**9.01**

Table 1. Performance comparison with state-of-the-art DiT-based approaches under various model sizes on ImageNet 256×256 for class-conditional image generation (batch size: 256).

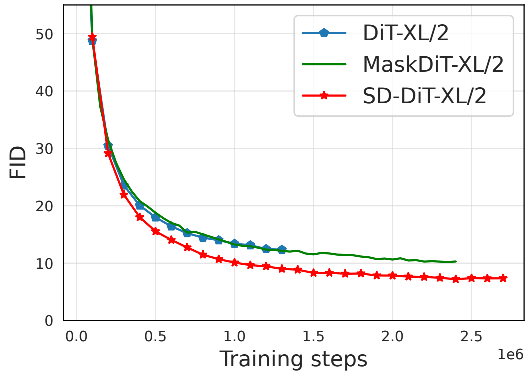


Figure 4. Comparison of convergence speed with SOTA DiT-based approaches in DiT-XL backbone (batch size: 256). The results of DiT and MaskDiT are directly cited from MaskDiT [81]. Our SD-DiT-XL/2 consistently outperforms DiT-XL/2 and MaskDiT-XL/2 across training steps, leading to better training convergence.

vantage of self-supervised discrimination knowledge distilling for mask modeling in Diffusion Transformer.

Comparison on Convergence Speed in Large Scale Backbone. Here we evaluate the convergence speed of our SD-DiT-XL/2 based on large-scale backbone. Fig. 4 illustrates the comparison of convergence speed by showing the FID scores in different training steps for our SD-DiT and various baselines. The batch size of each run is set as 256 for fair comparisons, and the maximum training step is 2400k. Note that the results of DiT and MaskDiT in different steps are directly copied from the reported results in MaskDiT [81]. As shown in Fig. 4, SD-DiT persistently reflects better training convergence than DiT and MaskDiT across the whole training steps. The detailed performance comparisons against MDT are listed in Tab. 1, where our SD-DiT (FID: 9.01) brings higher results than MDT (FID: 9.60) and MaskDiT (FID: 12.15) with 1300k training steps. It is worthy noting that SD-DiT trained with 1300k steps outperforms typical DiT with 7000k steps (FID: 9.01 *vs.* 9.62), achieving about 5× faster training progress. In addition, SD-DiT (1100k steps) achieves a comparable FID performance with MDT (1300k steps) (9.66 *vs.* 9.60). Such fast convergence again confirms the power of self-supervised discrimination for facilitating DiT training.

Method	Cost(Iter×BS)	FID↓	sFID↓	IS↑	Prec↑	Rec↑
VQGAN [19]	-	15.78	78.3	-	-	-
BigGAN-deep [5]	-	6.95	7.36	171.4	**0.87**	0.28
StyleGAN [61]	-	2.30	**4.02**	**265.12**	0.78	0.53
I-DDPM [47]	-	12.26	-	-	0.70	0.62
MaskGIT [10]	1387k×256	6.18	-	182.1	0.80	0.51
CDM [31]	-	**4.88**	-	158.71	-	-
ADM [17]	1980k×256	10.94	6.02	100.98	0.69	0.63
ADM-U [17]	—	7.49	5.13	127.49	0.72	0.63
LDM-8 [54]	4800k×64	15.51	-	79.03	0.65	0.63
LDM-4 [54]	178k×1200	10.56	-	103.49	0.71	0.62
MaskDiT-XL/2[81]	2000k×1024	5.69	10.34	177.99	0.74	0.60
DiT-XL/2 [49]	7000k×256	9.62	6.85	121.50	0.67	**0.67**
MDT-XL/2 [21]	2500k×256	7.41	4.95	121.22	0.72	0.64
SD-DiT-XL/2	2400k×256	7.21	5.17	144.68	0.72	0.61

Table 2. Performance comparison with state-of-the-art methods on ImageNet 256×256 for class-conditional image generation. Similar to most DiT-based approaches, here we report the results of our SD-DiT in DiT-XL backbone with 256 batch size, while MaskDiT reports results with the largest batch size (1024).

Method	FID
SD-DiT	53.72
w/o Discriminative Objective ($\mathcal{L}_{\mathrm{D}}$)	62.84
w/o Mask Strategy (mask ratio=0)	58.92

Table 3. Ablation studies on SD-DiT-S/2 with 400k training steps.

Comparison with State-of-the-Art Generative Methods. Tab. 2 summarizes the performance comparison against state-of-the-art generative methods. We strictly follow MDT [21] to list the cost comparison column as "Iter×Batchsize". We follow the most DiT-based approaches and report the results in DiT-XL backbone with larger training iterations (2400k). Generally, under the same batch size of 256, our SD-DiT-XL/2 achieves a better FID score than DiT-XL/2 and MDT-XL/2. Although MaskDiT-XL/2 obtains the best FID score among all DiT-based methods, it benefits from the extremely large batch size of 1024. A more fair comparison between our SD-DiT and MaskDiT can be referred to Fig. 4, where each run is trained with the same batch size (256). In that figure, SD-DiT-XL/2 leads to consistent performance boost against MaskDiT-XL/2, which clearly validates our proposal.

4.4. Ablation Study

We conduct ablation study to examine each component in SD-DiT. Considering that DiT training is computationally expensive, we adopt a lightweight setting for efficient evaluation: using small scale backbone (DiT-S) with 400k training steps, bs 256 and 50% mask ratio unless specified.

Effect of Discriminative Objective. Tab. 3 details the

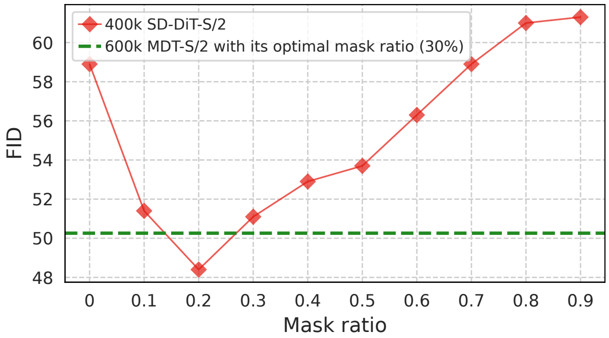


Figure 5. FID *vs.* mask ratio on SD-DiT-S/2 with 400k steps.

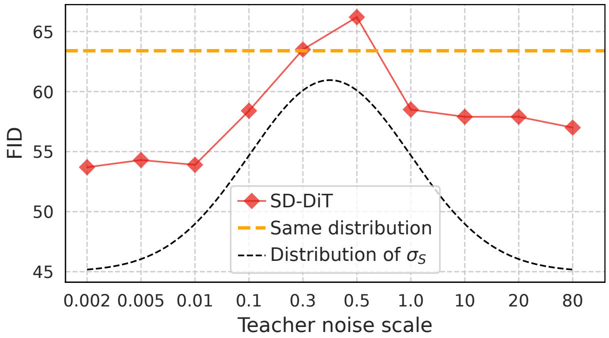


Figure 6. FID *vs.* teacher noise on SD-DiT-S/2 with 400k steps.

performances of ablated runs of our SD-DiT. Specifically, the first row shows the FID score (53.7) of our complete SD-DiT-S/2 with 50% mask ratio. Next, by removing discriminative objective ($\mathcal{L}_\mathrm{D}$) and the corresponding teacher branch from SD-DiT (2nd row), the generative performance drops by a large margin. This demonstrates the merit of our self-supervised discrimination tailored to Diffusion Transformer. In addition, when removing mask strategy of SD-DiT (3rd row), a clear performance drop is attained, which highlights the effectiveness of mask strategy that triggers the learning of intra-image contextual awareness [21, 81].

Effect of Mask Ratio. To further seek the sweet point of the balance between generative and discriminative task, we vary mask ratio from 0 to 1 and show the corresponding FID scores in Fig. 5. As shown in Fig. 5, the best performance of our SD-DiT is attained when the mask ratio is 20%, and thus we adopt this ratio practically in all experiments of Sec. 4.3. We additionally show the performance of MDT-S/2 trained with 600K steps under its optimal 30% mask ratio (the fixed green dashed line in Fig. 5, 50.3), which is inferior to our SD-DiT-S/2 with 400k steps (20% mask ratio, 48.4). Moreover, MaskDiT points out one interesting observation with regard to mask ratio: MaskDiT with 75% mask ratio achieves an extremely degraded FID score (121.16 of MaskDiT-XL/2). In other words, when 75% patches participate in the mask reconstruction task and only 25% local patches focus on the generative task, the generative ability of MaskDiT will be significantly weakened. This reveals the fuzzy relations between mask reconstruction and the generative task. Instead, in our SD-DiT-S/2, even when the mask ratio is increased to 90%, the corresponding FID score (61.0) is still higher than that of DiT-S/2 with 400k steps (68.40 in Tab. 1). These findings clearly verify that our design could alleviate the negative effect of fuzzy relations between mask modeling and generative task.

Effect of Noise of Teacher View. Recall that in our SD-DiT, the noise of student view is set as $\boldsymbol{x}_{\sigma_\mathrm{S}} = \boldsymbol{x}_0 + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathrm{S}^2\mathbf{I})$, $\sigma_\mathrm{S} \in [\sigma_\mathrm{min}, \sigma_\mathrm{max}]$ based on EDM formulation (Eq. (3)). Following EDM [34] and Consistency Model [64, 68], we set $\sigma_\mathrm{min} = 0.002$ and $\sigma_\mathrm{max} = 80$. Here we further test the effect of noise of teacher view. Specifically, we first set the noise of teacher view from the same distribution as student view, *i.e.*, $\boldsymbol{x}_{\sigma_\mathrm{T}} = \boldsymbol{x}_0 + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathrm{T}^2\mathbf{I})$, $\sigma_\mathrm{T} \in [\sigma_\mathrm{min}, \sigma_\mathrm{max}]$. As depicted in the yellow dashed line in Fig. 6, the corresponding FID (63.4) is somewhat unsatisfying. This result shows that teacher noise derived from the same distribution of student noise can not make the discriminative loss practical for generative task. Such observation aligns with InfoMin principle [69] in self-supervised learning: reducing the mutual information between two variant views can bring a good pre-train model learning with sufficient view-invariance. That's why we choose the fixed minimum noise as in Consistency Models [64, 68] for teacher view, *i.e.*, $\boldsymbol{x}_{\sigma_\mathrm{T}} = \boldsymbol{x}_0 + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathrm{min}^2\mathbf{I})$. In this way, the noise distribution of teacher view can be the closest one to the original data distribution ($\boldsymbol{x}_{\sigma_\mathrm{T}} \sim p_\mathrm{data}(\boldsymbol{x})$) and far away from student view. We empirically evaluate various teacher noise within $[\sigma_\mathrm{min}, \sigma_\mathrm{max}]$ (see the red curve in Fig. 6), and the fixed minimum noise (scale: 0.002) can get the best performance (53.7). Furthermore, we draw the approximate log-normal probability density distribution (PDF) of σ_S based on EDM (see the black dashed line in Fig. 6). When the fixed σ_T is set within the scale with high density (*e.g.*, 0.3 and 0.5 close to the mean of σ_S), the corresponding FID of SD-DiT drops drastically (*e.g.*, 66.2 when $\sigma_\mathrm{T} = 0.5$) and is even worse than the case of the same distribution (yellow dashed line). This again reveals that the noise scale of teacher view should be far away from the distribution of σ_S.

5. Conclusions

In this work, we propose a Diffusion Transformer architecture, namely SD-DiT, to facilitate the training process by unleashing the power of self-supervised discrimination. SD-DiT novelly frames mask modeling in a teacher-student manner to jointly execute discriminative and generative diffusion processes in a decoupled encoder-decoder structure. Such design nicely explores the mutual but also fuzzy relations between mask modeling and generative objective, leading to both effective and efficient DiT training. Experiments conducted on ImageNet validate the competitiveness of SD-DiT when compared to SOTA DiT-based approaches.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023. 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2

[3] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 3

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 3, 7

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 4, 5, 6

[10] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 1, 3, 7

[11] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3

[12] Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Controlstyle: Text-driven stylized image generation using diffusion priors. In *ACM Multimedia*, 2023. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 3

[15] Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. Control3d: Towards controllable text-to-3d generation. In *ACM Multimedia*, 2023. 1

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1, 3

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3, 6, 7

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3

[19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3, 7

[20] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[21] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *ICCV*, 2023. 1, 3, 4, 6, 7, 8

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 3

[23] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 1

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 3, 4, 5, 6

[26] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[27] 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. In *NIPS*, 2017. 6

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2

[30] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[31] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 7

[32] Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. *arXiv preprint arXiv:2212.11972*, 2022. 3

[33] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[34] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 2, 3, 8

[35] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019. 6

[36] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, 2023. 3

[37] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual Transformer Networks for Visual Recognition. *IEEE Trans. on PAMI*, 2022. 3

[38] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023. 3

[39] Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Jiebo Luo, and Tao Mei. Stand-Alone Inter-Frame Attention in Video Models. In *CVPR*, 2022. 3

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[41] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 2

[42] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2

[43] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *CVPR*, 2023. 1

[44] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2

[45] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. In *ICML*, 2021. 6

[46] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[47] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 7

[48] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1, 3, 6, 7

[50] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 6

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[52] 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 *ICML*, 2021. 3

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4, 6, 7

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[56] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 2

[58] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 2

[59] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 6

[60] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *ICLR*, 2017. 3

[61] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 2022. 7

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3

[64] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023. 4, 8

[65] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2, 3

[66] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020. 2

[67] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1, 2, 3

[68] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 4, 8

[69] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 4, 8

[70] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 3

[71] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NIPS*, 2017. 3

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 1, 3

[73] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary latent diffusion. In *CVPR*, 2023. 2

[74] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. *arXiv preprint arXiv:2304.03283*, 2023. 3

[75] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 3

[76] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, and Tao Mei. 3dstyle-diffusion: Pursuing fine-grained text-driven 3d stylization with 2d diffusion models. In *ACM Multimedia*, 2023. 1

[77] Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *arXiv preprint arXiv:2208.07791*, 2022. 3

[78] Ting Yao, Yehao Li, Yingwei Pan, Yu Wang, Xiao-Ping Zhang, and Tao Mei. Dual vision transformer. *IEEE Trans. on PAMI*, 2023. 3

[79] Ting Yao, Yehao Li, Yingwei Pan, and Tao Mei. Hiri-vit: Scaling vision transformer with high resolution inputs. *arXiv preprint arXiv:2403.11999*, 2024. 3

[80] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 1

[81] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 1, 3, 4, 5, 6, 7, 8

[82] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 3, 5, 6