Structure-Guided Adversarial Training of Diffusion Models

Ling Yang[*†]   Haotian Qian[*]   Zhilong Zhang   Jingwei Liu   Bin Cui[†]
Peking University

yangling0818@163.com, zzdmqht@gmail.com, {zzl2018math, bin.cui}@pku.edu

Abstract

Diffusion models have demonstrated exceptional efficacy in various generative applications. While existing models focus on minimizing a weighted sum of denoising score matching losses for data distribution modeling, their training primarily emphasizes instance-level optimization, overlooking valuable structural information within each mini-batch, indicative of pair-wise relationships among samples. To address this limitation, we introduce Structure-guided Adversarial training of Diffusion Models (SADM). In this pioneering approach, we compel the model to learn manifold structures between samples in each training batch. To ensure the model captures authentic manifold structures in the data distribution, we advocate adversarial training of the diffusion generator against a novel structure discriminator in a minimax game, distinguishing real manifold structures from the generated ones. SADM substantially outperforms existing methods in image generation and cross-domain fine-tuning tasks across 12 datasets, establishing a new state-of-the-art FID of 1.58 and 2.11 on ImageNet for class-conditional image generation at resolutions of 256×256 and 512×512, respectively.

1. Introduction

Diffusion models [20, 56–59, 73] have achieved remarkable generation quality in various tasks, including image generation [12, 50–52, 67, 74, 75, 79], audio synthesis [4, 33, 49], and interdisciplinary applications [21, 24, 71]. Starting from tractable noise distribution, diffusion models generate data by progressively removing noise. This involves the model learning to reverse a pre-defined diffusion process that sequentially introduces varying levels of noise to the data. The model is parameterized and undergoes training by optimizing the weighted sum of denoising score matching losses [20] for various noise levels [57], aiming to learn the recovery of clean images from corrupted images.

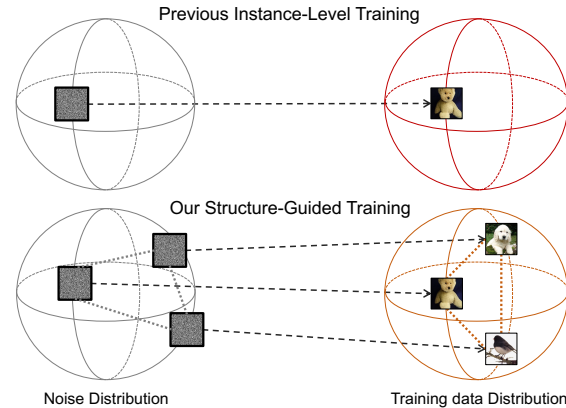[*]Contributed equally.
[†]Corresponding authors.

Figure 1. Comparison between previous instance-level training and our structure-guided training for diffusion models.

Aiming to maximally model the data distribution, recent works [28, 42, 60, 63, 72, 76] attempt to improve the precision of training diffusion models. For instance, many approaches design effective weighting schemes [8, 17, 30, 42, 60, 76] for maximum likelihood training or add an extra regularization term [10, 37, 46] to the denoising score loss. There are also some works to enhance the model expressiveness by incorporating other generative models, such as VAE [31, 51, 63], GAN [69], and Normalizing Flows [29, 43, 78]. However, their improved training of diffusion models primarily concentrates on instance-level optimization, overlooking the valuable structural information among batch samples. This oversight is significant, as the incorporation of structural details is crucial for aligning the learned distribution with the underlying data distribution.

To mitigate the challenges posed by existing instance-level training methods, we propose Structure-guided Adversarial training of Diffusion Models (SADM). In contrast to conventional instance-level training illustrated in Fig. 1, our approach guides diffusion training at a structural level. During batch training, the model is facilitated to learn manifold structures within batch samples, represented by pair-wise relationships in a low-dimensional feature space. To accurately learn real manifold structures in the data distribution, we introduce a novel *structure discriminator* that distinguishes genuine manifold structures from generated

FID: 4.59, Recall: 0.52   FID: **3.14**, Recall: **0.61**

ADM-G   ADM + **Structure-guided Adversarial Training**   Ground Truth
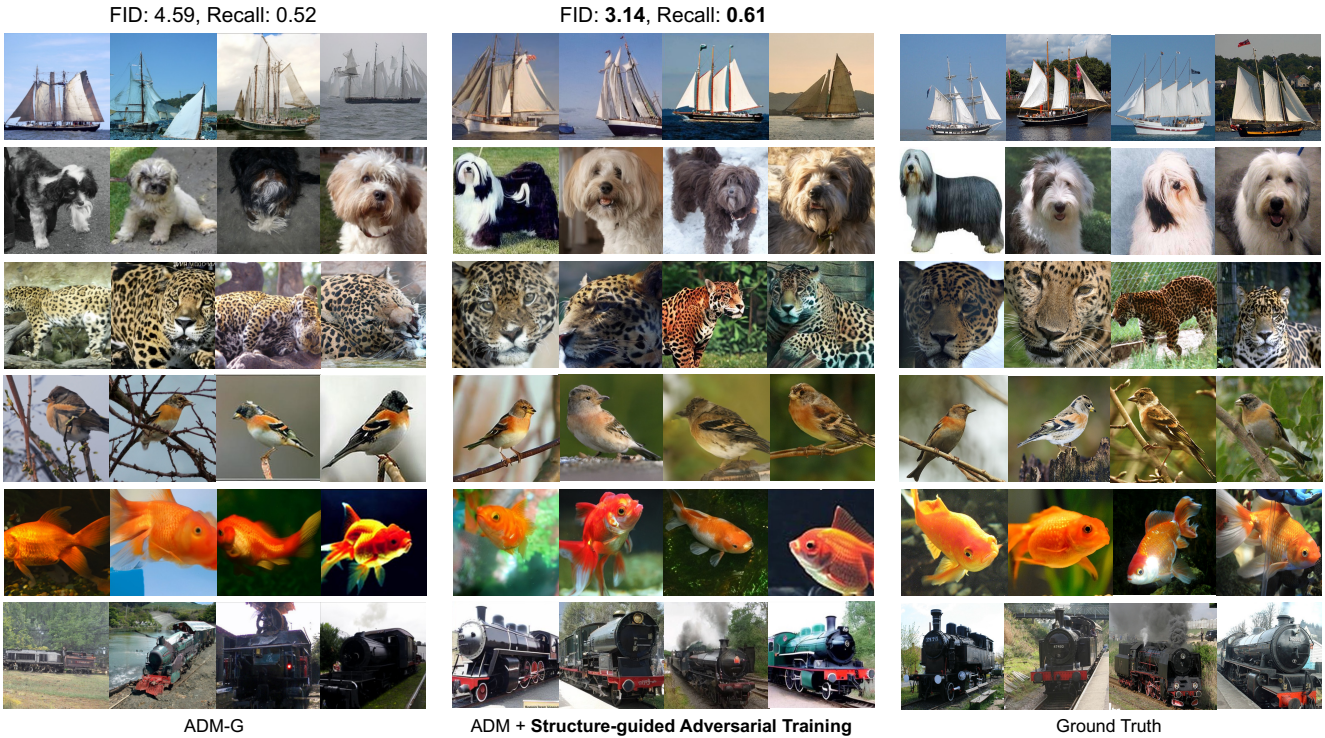
Figure 2. Generated samples on ImageNet 256×256 with (i) ADM with Classifier Guidance (ADM-G) [12], (ii) ADM optimized by our **Structure-guided Adversarial Training**, and (iii) real samples in ground truth classes. We can significantly improve diffusion models qualitatively and quantitatively, and our generated sample distribution is overally more similar to real sample distribution. See Appendix C for more synthesis samples of our SOTA model.

ones. For clarity, we alternatively refer to the diffusion models optimized through our structure-guided training as *joint sample diffusion*, a concept theoretically proven to enhance diffusion model optimization.

We assess the performance of our model across two pivotal tasks: image generation and cross-domain fine-tuning. The former involves training the diffusion model from scratch to evaluate its capability in capturing the entire data distribution. The latter leverages a pre-trained diffusion model, constructed on a large-scale source dataset, and fine-tunes it on a target dataset to assess transferability. Our extensive experiments consistently demonstrate that our approach significantly improves the model's ability to effectively capture the underlying data distribution (Fig. 2). SADM achieves state-of-the-art results across 12 image datasets, including ImageNet [11]. Furthermore, we observe its potential for facilitating rapid adaptation to new domains in cross-domain fine-tuning tasks. We summarize our contributions as follows:

- To the best of our knowledge, we are the first to propose **Structure-guided Adversarial Training** to optimize diffusion models from a structural perspective.
- We theoretically show that SADM is superior in capturing real data distribution, and can generalize to various image-

and latent-based diffusion architectures (e.g., DiT [47]).
- We substantially outperform existing methods on image generation and cross-domain fine-tuning tasks, achieving a new **state-of-the-art FID of 1.58 and 2.11** on ImageNet at resolutions of 256×256 and 512×512, respectively.

2. Related Work

In this work, we focus on **improving the training of diffusion models**. Here, we review previous related works and compare our SADM with them.

Modifying Training Objectives of Diffusion Models A line of research modifies training objectives to achieve state-of-the-art likelihood [8, 42, 60, 74]. Song et al. [60] propose likelihood weighting to enable approximate maximum likelihood training of score-based diffusion models [57–59] while ContextDiff [76] introduces an effective shifting scheme for facilitating the diffusion and training processes of diffusion probabilistic models [20, 56] to achieve improved sample quality with stable training. Lai et al. [37] and [10] introduce an extra regularization term to the denoising score loss to satisfy some properties of the diffusion process. However, these improvements mainly focus on sample-level optimization, neglecting the rich structural information within batch

samples, which is critical for aligning the learned distribution and data distribution. Hence, we enforce the model to maximally learn the manifold structures of samples.

Combining Additional Models for Diffusion Training
Another line of research incorporates other models to improve the stability and precision of diffusion training. For example, INDM [29] expands the linear diffusion to trainable nonlinear diffusion through a normalizing flow to improve the training curve of diffusion models. Jolicoeur-Martineau et al. [26], Kim et al. [28] improve diffusion models with adversarial learning while Xiao et al. [69] model each denoising step using a multimodal conditional GAN. LSGM [63] and LDM [51] conduct diffusion process in the semantic latent space obtained with a pre-trained VAE. Although these combinations strengthen the model expressiveness for capturing data distribution, they are still limited in modeling the underlying manifold structures within training samples. We propose a novel *structure discriminator* for adversarially learn the diffusion model from a structural perspective.

3. Preliminary

Diffusion Models   We consider diffusion models [20, 56, 57] specified in continuous time [4, 31, 59, 62]. Given samples $\mathbf{x}_0$ from a data distribution $q_0(\mathbf{x}_0)$, noise scheduling functions α_t, σ_t, a diffusion model has latent variables $\mathbf{x} = \{\mathbf{x}_t \mid t \in [0,1]\}$, and the forward process is defined with $q(\mathbf{x}_t|\mathbf{x}_0)$, a Gaussian process satisfying the following Markovian structure:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \qquad (1)$$

$$q(\mathbf{x}_t|\mathbf{x}_s) = \mathcal{N}(\mathbf{x}_t; (\alpha_t/\alpha_s)\mathbf{x}_s, \sigma_{t|s}^2 \mathbf{I}) \qquad (2)$$

where $0 \le s < t \le 1$ and $\sigma_{t|s}^2 = (1 - e^{\lambda_t - \lambda_s})\sigma_t^2$, and $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ denotes the log signal-to-noise-ratio [31]. The goal of the diffusion model is to denoise $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$ by estimating $\hat{\mathbf{x}}_\theta(\mathbf{x}_t) \approx \mathbf{x}_0$. We train this denoising model $\hat{\mathbf{x}}_\theta$ using a weighted mean squared error loss

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\left[w(\lambda_t)\|\hat{\mathbf{x}}_\theta(\mathbf{x}_t) - \mathbf{x}_0\|_2^2\right] \qquad (3)$$

over uniformly sampled times $t \in [0,1]$. This loss can be justified as a weighted variational lower bound on the data log likelihood under the diffusion model [31] or as a form of denoising score matching [57, 64]. $w(\lambda_t)$ is a pre-specified weighting function [31].

4. Proposed Method

We introduce the proposed SADM in detail (Fig. 3). In order to maximally learn the structural information between real samples, we propose structure-guided training of diffusion models in Sec. 4.1. Then we design a novel *structure*

discriminator for adversarially optimizing the training procedure from a structural perspective in Sec. 4.2. Finally, we alternatively interpret our structure-guided training of diffusion models as *joint sample diffusion* with theoretical analysis for bettwen understanding in Sec. 4.3.

4.1. Beyond Instance-Level Training

We first review the previous instance-level training methods for diffusion models. Generally, they train the diffusion model with a finite sample version of Eq. (3):

$$\mathcal{L}_t = w(\lambda_t)\frac{\sum_{i \in B} ||\mathbf{x}_0^i - \hat{\mathbf{x}}_\theta(\mathbf{x}_t^i)||^2}{|B|}, \qquad (4)$$

where $\mathbf{x}_0^i, \mathbf{x}_t^i$ denote the ith ground truth samples and generated samples in the mini-batch B at time step t. However, this objective function encourages the diffusion model to denoise by considering only the instance-level information, neglecting the group-level (structural) information in the mini-batch. Therefore, we enforce the sample predictions of the denoising network to maximally preserve the manifold structures between batch samples.

Structural Constraint in Manifold   More concretely, ground truth samples $\{\mathbf{x}_0^i\}_{i=1}^{|B|}$ are first projected from pixel space into embedding space using off-the-shelf pre-trained networks $\Psi(\cdot) : \mathbb{R}^{m \times n} \to \mathbb{R}^d$, *e.g.*, an Inception-V3 feature extractor pre-trained on ImageNet [61] (we conduct ablation study in Appendix B). Then, the pair-wise relationships $\mathcal{R}(\Psi(\mathbf{x}_0^i), \Psi(\mathbf{x}_0^j))$ are calculated within batch samples $\{\mathbf{x}_0^i\}_{i=1}^{|B|}$, which contain rich structural information in a low-dimensional manifold. This structural information can be expressed with an affinity matrix $\mathcal{M}(\{\mathbf{x}_0^i\}_{i=1}^{|B|})$ defined as:

$$\begin{bmatrix} \mathcal{R}(\Psi_1, \Psi_1) & \mathcal{R}(\Psi_1, \Psi_2) & \cdots & \mathcal{R}(\Psi_1, \Psi_{|B|}) \\ \mathcal{R}(\Psi_2, \Psi_1) & \mathcal{R}(\Psi_2, \Psi_2) & \cdots & \mathcal{R}(\Psi_2, \Psi_{|B|}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{R}(\Psi_{|B|}, \Psi_1) & \mathcal{R}(\Psi_{|B|}, \Psi_2) & \cdots & \mathcal{R}(\Psi_{|B|}, \Psi_{|B|}) \end{bmatrix} \qquad (5)$$

where Ψ_i is a short-hand for $\Psi(\mathbf{x}_0^i)$, and $\mathcal{R}(\cdot, \cdot)$ denotes the relational function, such as Euclidean distance. Kindly note that many off-the-shelf pre-trained networks are open-source and only work well on clean images, and thus fail to provide meaningful embeddings when the input is noisy. Therefore, we use the predicted clean samples $\hat{\mathbf{x}}_0^i$ for computing affinity matrix, and regularize the denoising network to minimize the structural distance between $\mathcal{M}(\{\mathbf{x}_0^i\}_{i=1}^{|B|})$ and $\mathcal{M}(\{\hat{\mathbf{x}}_0^i\}_{i=1}^{|B|})$. Adding this structural constraint into Eq. (4), the training objective for denoising network is:

$$\mathcal{L}_t = \frac{\sum_{i \in B} ||\mathbf{x}_0^i - \hat{\mathbf{x}}_\theta(\mathbf{x}_t^i)||^2}{|B|} \\ + \mathcal{D}(\mathcal{M}(\{\mathbf{x}_0^i\}_{i=1}^{|B|}), \mathcal{M}(\{\hat{\mathbf{x}}_0^i\}_{i=1}^{|B|})) \qquad (6)$$
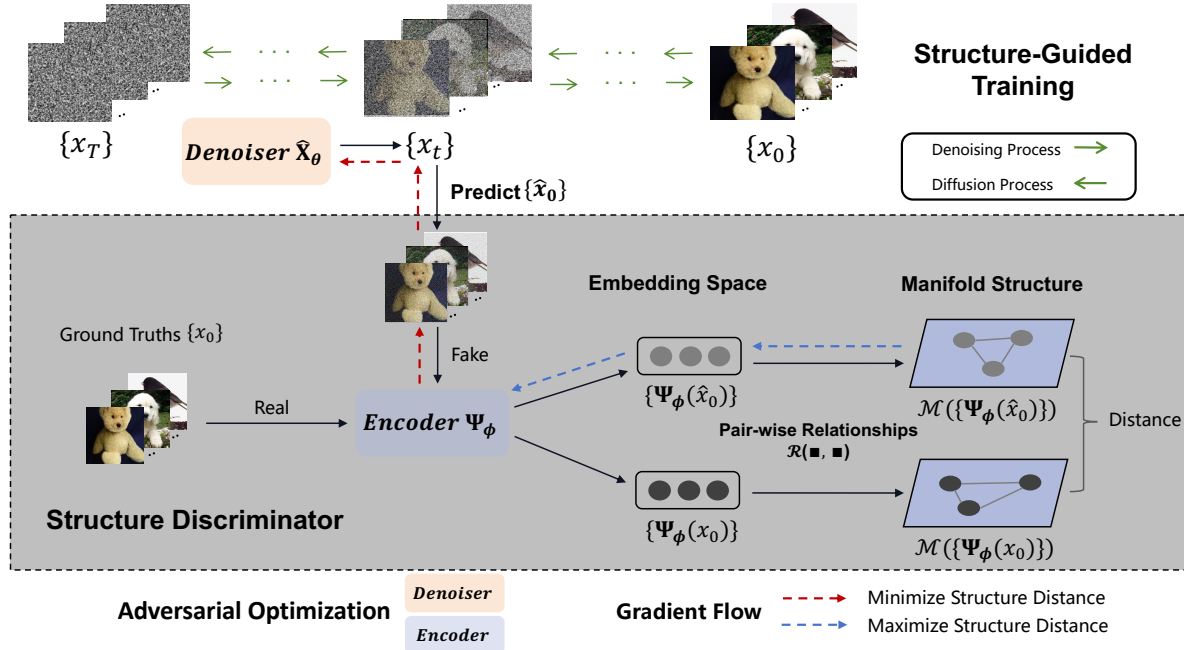
Figure 3. **Overview of SADM.** We minimize the structural distance between the generated samples (fake) and ground truth samples (real) in the manifold space for optimizing the denoiser, and maximize their structural distance for optimizing the encoder in *structure discriminator*. The denoiser and the structure discriminator are adversarially trained.

where $\mathcal{D}(\cdot, \cdot)$ denotes the distance metric between ground truth and predicted affinity matrices. In this way, the diffusion generator (denoiser) is optimized not only to make correct prediction for each instance, but also to preserve the manifold structures of batch samples.

4.2. Structure-Guided Adversarial Training

As demonstrated above, we aim to maximally learn data distribution by aligning the manifold structures of denoiser's predicted samples to those of ground truth samples in each training batch. At every training iteration, the manifold structures can be diverse and the denoiser tend to merely focus on some easy-to-learn structures, ultimately leading to trivial solutions that fail to capture the whole data distribution. In order to mitigate this problem and improve the expressiveness of denoiser, we adversarially learn the denoising network against a *structure discriminator* in a minimax game (Fig. 3), which is trained to distinguish the manifold structures between the real and generated batch samples.

Structure Discriminator Normally, the discriminator in adversarial learning [15] would output the discrete value (0 or 1) to determine whether the input is real or fake. However, such discrete discriminator can not be applied in distinguishing manifold structures, because the generated and real sample sets would share similar pair-wise sample relations despite their different feature spaces. Thus, instead of learning a classification-based discriminator, we design a

novel comparison-based structure discriminator to address this issue.

Specifically, as illustrated in Fig. 3, the structure discriminator consists of the aforementioned neural network Ψ_ϕ with trainable parameters ϕ and the distance measure function $\mathcal{D}(\cdot, \cdot)$ that outputs continuous value. It projects both real and generated samples into the embedding space, then the structure discriminator is trained against the denoiser for finding a better manifold (or embedding space) to distinguish real sample set from generated set by maximizing their structural distance:

$$\max_\phi \frac{\eta_t \sum_{i,j \in B} \mathcal{D}\big(\mathcal{R}(\Psi_\phi(\mathbf{x}_0^i), \Psi_\phi(\mathbf{x}_0^j)), \mathcal{R}(\Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^i), \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^j))\big),}{|B|^2}.$$
(7)

where we choose $\eta_t = \frac{1}{t}$ as a time-dependent weighting factor. We can use simple measurements for $\mathcal{D}(\cdot, \cdot)$ (L_2 distance) and $\mathcal{R}(\cdot, \cdot)$ (cosine similarity), as the critical semantic information has been already encoded by Ψ_ϕ. Conversely, the denoiser is adversarially optimized for generating more realistic sample set to fool the structure discriminator by minimizing the structural distance.

Final Optimization Objective The final training objective of our SADM consists of normal denoising score matching loss (Eq. (3)) and adversarial structural distance loss (Eq. (7))

at timestep t, which can be written as:

$$\mathcal{L}_t(\theta) = \frac{w(\lambda_t) \sum_{i \in B} \|\hat{\mathbf{x}}_\theta(\mathbf{x}_t^i) - \mathbf{x}_0^i\|_2^2}{|B|}$$
$$+ \max_\phi \frac{\sum_{i,j \in B} \eta_t \|\Psi_\phi(\mathbf{x}_0^i)^T \Psi_\phi(\mathbf{x}_0^j) - \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^i)^T \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^j)\|_2^2}{|B|^2}.$$
(8)

In training, we iteratively optimize the feature extractor and the denoising network in Eq. (8). In an iteration, we first freeze θ and update ϕ by ascending along its gradient

$$\nabla_\phi \frac{\sum_{i,j \in B} \eta_t \|\Psi_\phi(\mathbf{x}_0^i)^T \Psi_\phi(\mathbf{x}_0^j) - \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^i)^T \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^j)\|_2^2}{|B|^2},$$
(9)

then we freeze ϕ and update θ by descending along its gradient

$$\nabla_\theta \frac{\sum_{i,j \in B} \eta_t \|\Psi_\phi(\mathbf{x}_0^i)^T \Psi_\phi(\mathbf{x}_0^j) - \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^i)^T \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^j)\|_2^2}{|B|^2}$$
$$+ \nabla_\theta \frac{w(\lambda_t) \sum_{i \in B} \|\hat{\mathbf{x}}_\theta(\mathbf{x}_t^i) - \mathbf{x}_0^i\|_2^2}{|B|}.$$
(10)

The proposed training algorithm is presented in Algorithm 1.

Generalizing to Latent Diffusion   Our proposed training algorithm applies not only to image diffusion but also to latent diffusion, such as LDM [51] and LSGM [63]. In this case, the intermediate results $\mathbf{x}_t$ are latent codes rather than images. We can use the latent decoder (*e.g.*, VAE decoder [32]) to project the generated latent codes to images and then use the same algorithm in the image domain.

4.3. Interpreting as Joint Sample Diffusion

Relation-Conditioned Diffusion Process   For better understanding of our proposed structure-guided training, we interpret it as a joint sample diffusion model that simultaneously perturbs and denoises a set of samples conditioned on the relation variable. Formally, let $\mathbf{y}_0 = (\mathbf{x}_0^i, \mathbf{x}_0^j)$, where $\mathbf{x}_0^i, \mathbf{x}_0^j$ are independent random variables sampled from ground truth distribution q_0. And the relation variable that encodes the structure information is defined as $\mathcal{R} = \mathcal{R}(\mathbf{x}_0^i, \mathbf{x}_0^j) + \gamma\epsilon$, where $\gamma\epsilon$ denote a small gaussian noise added to the relation to avoid degenerated distribution. Then the forward diffusion jointly perturbs $\mathbf{y}_0$, conditioned on $\mathcal{R}$:

$$q(\mathbf{y}_t|\mathbf{y}_0, \mathcal{R}) = \mathcal{N}(\mathbf{y}_t; \alpha_t \mathbf{y}_0, \sigma_t^2 \mathbf{I}).$$
(11)

To reverse the diffusion process, we need to predict $\mathbf{y}_0$, or equivalently, learn the conditional score function $\nabla_y \log q_t(y_t|\mathcal{R})$ [12, 19]. This formulation allows us to utilize the auxiliary structural information $\mathcal{R}$ in the sampling process, which usually leads to better performance [13, 55].

Algorithm 1: SADM, our proposed algorithm.

input : $\{\alpha_t\}_{t \in [0,1]}, \{\sigma_t\}_{t \in [0,1]}$ the noise schedule, $|B|$ the batch size, $w(\lambda_t), \eta_t$ the scaling factors for denoising score matching loss and adversarial loss, pre-trained encoder Ψ_ϕ for structure discriminator.

Initialize denoising network parameters θ;

while θ *has not converged* **do**

> Sample a minibatch $\{\mathbf{x}_0^i\}_{i=1}^{|B|} \sim q_0$ and $\{\epsilon^j\}_{j=1}^{|B|} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
> Sample $t \sim U[0,1]$
> Sample $\{\mathbf{x}_t^i = \alpha_t \mathbf{x}_0^i + \sigma_t \epsilon^i\}_{i=1}^{|B|}$,
> Freeze ϕ and update θ with its gradient
> $\nabla_\theta \frac{\sum_{i,j \in B} \eta_t \|\Psi_\phi(\mathbf{x}_0^i)^T \Psi_\phi(\mathbf{x}_0^j) - \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^i)^T \Psi_\phi(\hat{\mathbf{x}}_{\theta,t}^j)\|_2^2}{|B|^2}$
> $+ \nabla_\theta \frac{w(\lambda_t) \sum_{i \in B} \|\hat{\mathbf{x}}_\theta(\mathbf{x}_t^i) - \mathbf{x}_0^i\|_2^2}{|B|}$

Adversarially train denoising network and structure discriminator by iteratively updating their parameters θ, ϕ according to Eqs. (9) and (10).

output : Denoising network θ.

Learning Conditional Score   To approximate the conditional score function, first we decompose it with Bayes' rule,

$$\nabla_y \log q_t(y_t|\mathcal{R}) = \nabla_y \log q_t(y_t) + \nabla_y \log q_t(\mathcal{R}|y_t)$$
$$= \Sigma_{s=i,j} \nabla_x \log q_t(x_t^s) + \nabla_y \log q_t(\mathcal{R}|y_t).$$
(12)

To approximate the first term on the right side of Eq. (12), we only need to learn $\hat{\mathbf{x}}_{\theta,t}^i, \hat{\mathbf{x}}_{\theta,t}^j$, as in standard diffusion models [59]. However, the second term is intractable since $q_t(\mathcal{R}|y_t)$ involves the intractable posterior $q(\mathbf{y}_0|\mathbf{y}_t)$. Note that

$$q_t(\mathcal{R}|y_t) = \int g(\mathcal{R}|\mathbf{y}_0) q(\mathbf{y}_0|\mathbf{y}_t) d\mathbf{y}_0$$
$$= \mathbb{E}_{\mathbf{y}_0} [g(\mathcal{R}|\mathbf{y}_0)|\mathbf{y}_t],$$
(13)

where $g(\mathcal{R}|\mathbf{y}_0)$ is the density function of Gaussian distribution with mean $\mathcal{R}(\mathbf{x}_0^i, \mathbf{x}_0^j)$ and variance γ^2, by the definition of $\mathcal{R}$. As a result, we can use $g(\mathcal{R}|\hat{\mathbf{y}}_{0,t}) = g(\mathcal{R}|(\hat{\mathbf{x}}_{\theta,t}^i, \hat{\mathbf{x}}_{\theta,t}^j))$ to approximate $q_t(\mathcal{R}|y_t)$, and train it with L_2 loss:

$$\mathbb{E}_{\mathbf{y}_t} \|\mathbb{E}_{\mathbf{y}_0} [g(\mathcal{R}|\mathbf{y}_0)|\mathbf{y}_t] - g(\mathcal{R}|\hat{\mathbf{y}}_0(\mathbf{y}_t))\|_2^2$$
$$= \mathbb{E}_{\mathbf{y}_t} \|\mathbb{E}_{\mathbf{y}_0} [g(\mathcal{R}|\mathbf{y}_0) - g(\mathcal{R}|\hat{\mathbf{y}}_0(\mathbf{y}_t))|\mathbf{y}_t]\|_2^2$$
$$\leq \mathbb{E}_{\mathbf{y}_0, \mathbf{y}_t} \|g(\mathcal{R}|\mathbf{y}_0) - g(\mathcal{R}|\hat{\mathbf{y}}_0(\mathbf{y}_t))\|_2^2 \quad \text{(\textit{Jensen Inequality})}$$
$$\leq w(\gamma) \mathbb{E}_{\mathbf{y}_0, \mathbf{y}_t} \|\mathcal{R}(\mathbf{x}_0^i, \mathbf{x}_0^j) - \mathcal{R}(\hat{\mathbf{x}}_{\theta,t}^i, \hat{\mathbf{x}}_{\theta,t}^j)\|_2^2 = \mathcal{L}_t^{\text{structure}},$$
(14)

where $w(\gamma)$ is a weighting scalar that depends only on γ. The objective function is the sum of the denoising score

Table 1. Quantitative results for class-conditional generation on ImageNet 256×256 and 512×512.

Model	ImageNet 256×256				ImageNet 512×512			
	FID ↓	IS ↑	Precision ↑	Recall ↑	FID ↓	IS ↑	Precision ↑	Recall ↑
BigGAN-deep [2]	6.95	171.40	**0.87**	0.28	8.43	177.90	**0.88**	0.29
StyleGAN-XL [54]	2.30	265.12	0.78	0.53	2.41	267.75	0.77	0.52
ADM-G, ADM-U [12]	3.94	215.84	0.83	0.53	3.85	221.72	0.84	0.53
LDM-4-G [51]	3.60	247.67	**0.87**	0.48	-	-	-	-
RIN+NoiseSchedule [6]	3.52	186.20	-	-	3.95	216.00	-	-
SimpleDiffusion [22]	2.44	256.30	-	-	3.02	248.70	-	-
DiT-G++ [28]	1.83	281.53	0.78	0.64	-	-	-	-
MDT-G [14]	1.79	283.01	0.81	0.61	-	-	-	-
DiT-XL/2-G [47]	2.27	278.24	0.83	0.57	3.04	240.82	0.84	0.54
DiT-SADM (Ours)	**1.58**	**298.46**	**0.86**	**0.66**	**2.11**	**251.82**	**0.87**	**0.63**

matching objective and $\mathcal{L}_t^{\text{structure}}$:

$$\mathbb{E}_{\mathbf{x}_0^i, \boldsymbol{\epsilon}^i, t} w(\lambda_t) \|\hat{\mathbf{x}}_\theta(\mathbf{x}_t^i) - \mathbf{x}_0^i\|_2^2 + \mathbb{E}_{\mathbf{x}_0^j, \boldsymbol{\epsilon}^j, t} w(\lambda_t) \|\hat{\mathbf{x}}_\theta(\mathbf{x}_t^j) - \mathbf{x}_0^j\|_2^2$$
$$+ \mathbb{E}_{\mathbf{x}_0^i, \boldsymbol{\epsilon}^i, \mathbf{x}_0^j, \boldsymbol{\epsilon}^j, t} \left[\mathcal{L}_t^{\text{structure}} \right] , \quad (15)$$

which is a variational upper bound of negative log likelihood of the joint sample. Our objective in Eq. (6) can be viewed as a finite-sample version of Eq. (15), which trains the conditional diffusion model to utilize the structural information.

5. Experiments

5.1. Image Generation

Experiment Setup We experiment on CIFAR-10 [35], CelebA/FFHQ 64x64 [41], and ImageNet 256x256 [11]. We utilize our SADM to facilitate the training of the diffusion backbones from Karras et al. [27], Vahdat et al. [63] on CIFAR-10 and FFHQ, from Kim et al. [30] on CelebA, and from Peebles and Xie [47] (DiT, Diffusion Transformer) on ImageNet.

Evaluation Metrics We use Frechet Inception Distance (FID) [18] as the primary metric for capturing both quality and diversity due to its alignment with human judgement. We follow the evaluation procedure of ADM [12] for fair comparisons. For completeness, we also use Inception Score (IS) [53], Precision and Recall [36] as the main metrics for measuring diversity and distribution coverage.

Implementation Details We train the denoising network from scratch and use the feature extractor of Inception-V3 [61] pre-trained on ImageNet for initializing the encoder of our structure discriminator. At the begining of training, we freeze the pre-trained discriminator encoder and train the denoiser with structure-guided training objective in Eq. (6) until convergence, then we adversarially tune the denoiser and

Table 2. Performance on CIFAR-10.

Model	Diffusion Space	NFE↓	Unconditional		Conditional
			NLL↓	FID↓	FID↓
VDM [31]	Data	1000	2.49	7.41	-
DDPM [20]	Data	1000	3.75	3.17	-
iDDPM [44]	Data	1000	3.37	2.90	-
Soft Truncation [30]	Data	2000	2.91	2.47	-
INDM [29]	Latent	2000	3.09	2.28	-
CLD-SGM [13]	Data	312	3.31	2.25	-
NCSN++ [59]	Data	2000	3.45	2.20	-
LSGM [63]	Latent	138	3.43	2.10	-
NCSN++-G [3]	Data	2000	-	-	2.25
EDM [27]	Data	**35**	2.60	1.97	1.79
LSGM-G++ [28]	Latent	138	3.42	1.94	-
EDM-G++ [28]	Data	**35**	2.55	1.77	1.64
SADM	Latent	138	2.51	1.78	1.73
SADM	Data	**35**	**2.28**	**1.54**	**1.47**

encoder with the objective in Eq. (8) for 3 or 4 rounds (500k steps) until they achieve a balance. This training paradigm keeps the same for unconditional and class-conditional generation tasks, and can be easily generalized to score-based diffusion models [59, 63] by adding our structural constraint into the final objective functions.

Main Results Our SADM achieves new state-of-the-art FIDs on all datasets including CIFAR-10, CelebA, FFHQ, and ImageNet. On ImageNet 256×256 and 512×512, we consistently achieve SOTA FIDs of 1.43 and 2.18 for class-conditional generation as illustrated in Tab. 1. Notably, We significantly improve the generation performance of DiT and outperform the previous best FID of MDT [14] solely through improved training algorithm without increasing the model complexity and inference time. From Tab. 2, we find that our SADM works well for both image diffusion (based on EDM) and latent diffusion (based on LSGM). In experiments, for fair comparisons, we use the same hyperpa-
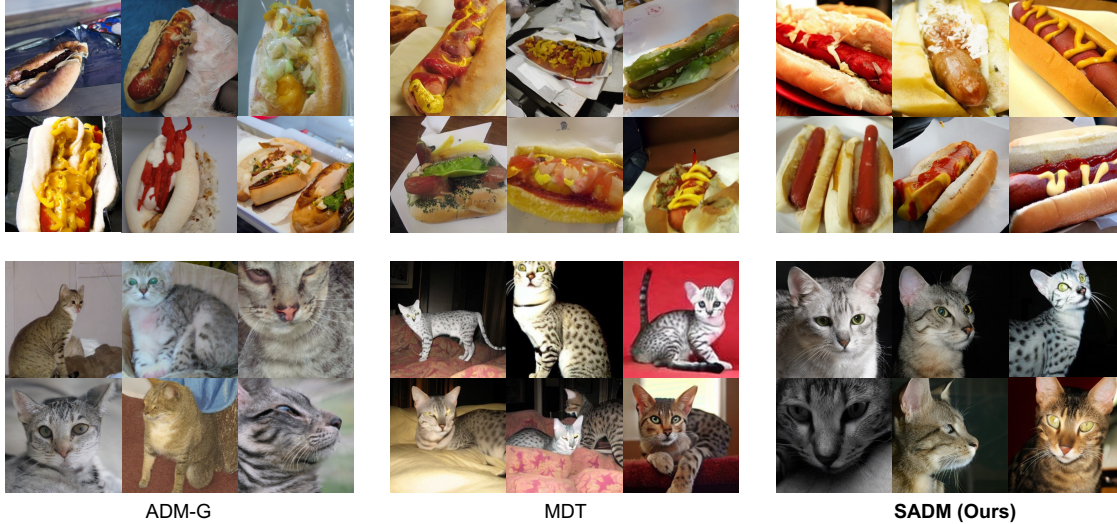
Figure 4. Qualitative comparion with ADM-G [12] and previous SOTA method MDT [14]. Our SADM can synthesize more realistic and high-quality samples while maintaining satisfying diversity.

Table 3. FID performance on CelebA/FFHQ 64×64.

Model	Diffusion Space	NFE↓	CelebA	FFHQ
DDPM++ [59]	Data	131	2.32	-
Soft Truncation [30]	Data	131	1.90	-
Soft Diffusion [9]	Data	300	1.85	-
INDM [29]	Latent	132	1.75	-
EDM [27]	Data	79	-	2.39
Soft Truncation-G++ [28]	Data	131	1.34	-
EDM-G++ [28]	Data	71	-	1.98
SADM	Latent	131	**1.28**	**1.85**
SADM	Data	71	**1.16**	**1.71**

rameters as EDM and LSGM to evaluate the effectiveness of our proposed training algorithm. And we also achieve significant performance improvement on facial datasets as demonstrated in Tab. 3. For qualitative results, we compare our SADM with ADM-G [12] and previous SOTA method MDT [14] in Fig. 4. These remarkable results demonstrate our SADM has the potential for generalizing to arbitrary diffusion architectures and can better learn the whole data distribution.

5.2. Cross-Domain Fine-Tuning

Experiment Setup We conduct cross-domain fine-tuning tasks on diffusion image generation for evaluating the transferability of proposed model, where we pre-train a diffusion model in source domain and adapt it to a target domain by fine-tuning. Following Xie et al. [70], we use ImageNet as source dataset, and choose eight commonly-used fine-grained datasets as target datasets: Food101 [1], SUN397 [68], DF-20M mini [48], Caltech101 [16], CUB-200-2011 [65], ArtBench-10 [40], Oxford Flowers [45] and Stanford

Cars [34]. More details about datasets are in Appendix A.

Implementation Details For fair comparison, we follow Xie et al. [70] to set all the hyper-parameters in both pre-training and fine-tuning stages, and use Diffusion Transformer (DiT) [47] as the diffusion backbone. We pretrain DiT on ImageNet 256×256 with a learning rate of 0.0001 using DDPM objective. For target datasets, we fine-tune the pre-trained DiT with 24k structure-guide training steps and 4k adversarial training steps. We experiment with two fine-tuning settings, *full* and *parameter-efficient*, for comprehensive evaluations. In parameter-efficient setting, following Xie et al. [70], we freeze most of parameters in the pre-trained diffusion model and fine-tune the bias term, normalization, and class condition module.

Main Results We achieve SOTA performance on all datasets in fine-tuning tasks as illustrated in Tab. 4. Remarkably, we significantly surpass DDPM in full fine-tuning and outperform DiffFit in parameter-efficient fine-tuning. The results sufficiently demonstrate our superior capability of capturing the whole data distribution, which enables better adaptation to new domains. Among all datasets, we achieve the best improvement over other methods on ArtBench-10 which has distinct distribution from ImageNet, demonstrating the out-of-distribution generalization ability of our SADM. More qualitative results are in Appendix C.

5.3. Model Analysis

Heatmap Analysis To evaluate the ability to capture data distribution, we perform heatmap analysis in Fig. 5, where we provide DDPM and our SADM with 8 randomly-selected

Table 4. FID performance comparisons on 8 downstream datasets, all the models are pretrained on ImageNet 256×256.

Dataset / Method	Food	SUN	DF-20M	Caltech	CUB-Bird	ArtBench	Oxford Flowers	Standard Cars	Average FID
AdaptFormer [5]	13.67	11.47	22.38	35.76	7.73	38.43	21.24	10.73	20.17
BitFit [77]	9.17	9.11	17.78	34.21	8.81	24.53	20.31	10.64	16.82
VPT [25]	18.47	14.54	32.89	42.78	17.29	40.74	25.59	22.12	26.80
LoRA [23]	33.75	32.53	120.25	86.05	56.03	80.99	164.13	76.24	81.25
DiffFit [70]	6.96	8.55	17.35	33.84	5.48	20.87	20.18	9.90	15.39
Full Fine-tuning with DDPM	10.46	7.96	17.26	35.25	5.68	25.31	21.05	9.79	16.59
SADM (parameter-efficient)	**5.74**	7.92	16.58	**32.03**	5.04	**18.23**	19.37	9.26	14.27
SADM (full)	6.20	**7.35**	**15.12**	32.86	**4.69**	19.84	**18.18**	**8.93**	**14.15**

noisy images in test batch and visualize the correlations between their denoised outputs. We observe that compared to DDPM, the overall heatmap pattern of our SADM is more closer to that of label affinity. The phenomenon demonstrates our SADM can precisely learn the manifold structures within real data samples.
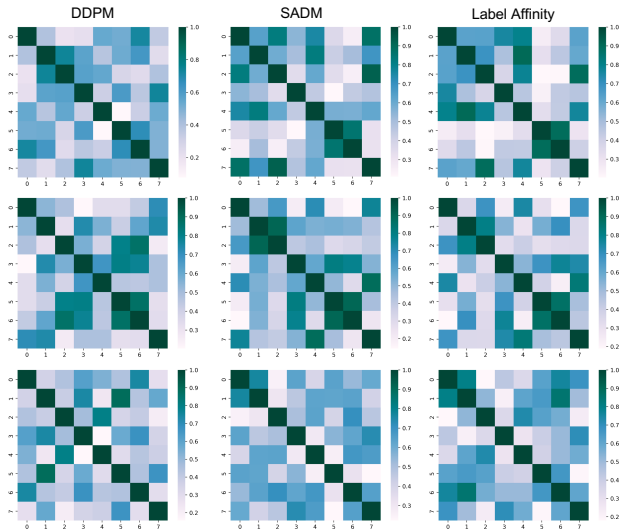

Figure 5. Heatmap visualization with 8 denoised samples.

Ablation Study We conduct ablation study to validate the effectiveness of our algorithm in Tab. 5. Here, we base on DDPM [20] architecture, and progressively add our model components (structural guidance and structure discriminator) into it for evaluating FID score on three datasets. We observe that each component can consistently improve the DDPM on all datasets, and the performance improvement of our structural guidance is more significant. The results fully demonstrate the effectiveness of our algorithm. More ablation studies about our model are in Appendix B.

Contributing to Better Convergence To investigate the contribution of our structure-guided training to model convergence, we plot the training curve in Fig. 6. We conclude that compared to previous SOTA methods DiT [47] and MDT

Table 5. Ablation study with FID performance. SG denotes structural guidance, SAT denotes SG+Structure Discriminator.

Model / Dataset	DDPM	+ Our SG	+ Our SAT
ImageNet	4.59	3.57	3.14
CIFAR-10	3.17	2.64	2.33
CelebA	2.32	1.82	1.64

[14], the proposed structure-guided training enables faster and better model convergence because we optimize the diffusion models from a structural perspective, which essentially contributes to capturing the whole data distribution.
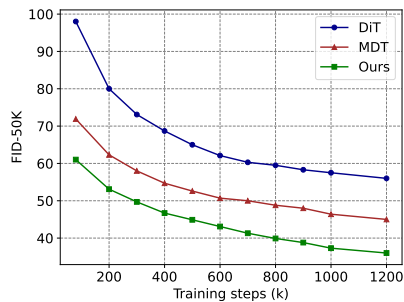

Figure 6. Comparison with SOTA methods on model convergence.

6. Conclusion

We propose structure-guided adversarial training for optimizing diffusion models from a structural perspective. The proposed training algorithm can easily generalize to both image and latent diffusion models, and consistently improve existing diffusion models with theoretical derivations and empirical results. We achieve new SOTA performance on image generation and cross-domain fine-tuning tasks across 12 image datasets. For future work, we will extend our method to more challenging diffusion-based applications (e.g., text-to-image/video generation).

Acknowledgement

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 7, 12

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 6

[3] Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, and Chun-Yi Lee. Denoising likelihood score matching for conditional score-based data generation. In *International Conference on Learning Representations*, 2022. 6

[4] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2020. 1, 3

[5] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv*, 2022. 8

[6] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 6

[7] Gong Cheng, Pujian Lai, Decheng Gao, and Junwei Han. Class attention network for image recognition. *Science China Information Sciences*, 66(3):132105, 2023. 13

[8] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, pages 11472–11481, 2022. 1, 2

[9] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions. *arXiv preprint arXiv:2209.05442*, 2022. 7

[10] Giannis Daras, Yuval Dagan, Alexandros G Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *arXiv preprint arXiv:2302.09057*, 2023. 1, 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 1, 2, 5, 6, 7, 14, 15, 16, 17, 18

[13] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations*, 2022. 5, 6

[14] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 6, 7, 8

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. 4

[16] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 7, 12

[17] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. *arXiv preprint arXiv:2303.09556*, 2023. 1

[18] 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. *NeurIPS*, 30, 2017. 6

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

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1, 2, 3, 6, 8

[21] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022. 1

[22] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 6

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arxiv*, 2021. 8

[24] Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin CUI, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[25] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 8

[26] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Ioannis Mitliagkas, and Remi Tachet des Combes. Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations*, 2020. 3

[27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. 2022. 6, 7

[28] Dongjun Kim, Yeongmin Kim, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022. 1, 3, 6, 7

[29] Dongjun Kim, Byeonghu Na, Se Jung Kwon, Dongsoo Lee, Wanmo Kang, and Il-chul Moon. Maximum likelihood training of implicit nonlinear diffusion model. *Advances in Neural Information Processing Systems*, 35:32270–32284, 2022. 1, 3, 6, 7

[30] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pages 11201–11228. PMLR, 2022. 1, 6, 7

[31] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *NeurIPS*, 34:21696–21707, 2021. 1, 3, 6

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[33] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020. 1

[34] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 7, 13

[35] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[36] 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

[37] Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Regularizing score-based models with score fokker-planck equations. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. 1, 2

[38] Zhi Lei, Guixian Zhang, Lijuan Wu, Kui Zhang, and Rongjiao Liang. A multi-level mesh mutual attention model for visual question answering. *Data Science and Engineering*, 7(4):339–353, 2022. 13

[39] Kun Li, Dan Guo, and Meng Wang. Vigt: proposal-free video grounding with a learnable token in the transformer. *Science China Information Sciences*, 66(10):202102, 2023. 13

[40] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv*, 2022. 7, 12

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 6

[42] Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International Conference on Machine Learning*, pages 14429–14460. PMLR, 2022. 1, 2

[43] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pages 2837–2845, 2021. 1

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning, ICML*, 2021. 6

[45] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 7, 13

[46] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *International Conference on Machine Learning*, pages 26245–26265. PMLR, 2023. 1

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 6, 7, 8

[48] Lukáš Picek, Milan Šulc, Jiří Matas, Thomas S Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020-not just another image recognition dataset. In *WACV*, 2022. 7, 12

[49] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021. 1

[50] 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

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3, 5, 6

[52] 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. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[53] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. 6

[54] Axel Sauer, Katja Schwarz, and Andreas Geiger. Styleganxl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*, pages 1–10, 2022. 6

[55] Raghav Singhal, Mark Goldstein, and Rajesh Ranganath. Where to diffuse, how to diffuse, and how to get back: Automated learning for multivariate diffusions. In *The Eleventh International Conference on Learning Representations*, 2022. 5

[56] 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, 2015. 1, 2, 3

[57] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 1, 2, 3

[58] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, pages 12438–12448, 2020.

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 5, 6, 7

[60] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021. 1, 2

[61] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 3, 6

[62] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019. 3

[63] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 1, 3, 5, 6

[64] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. 3

[65] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 7, 12

[66] Meng Wang, Yinghui Shi, Han Yang, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. Probing the impacts of visual context in multimodal entity alignment. *Data Science and Engineering*, 8(2):124–134, 2023. 13

[67] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022. 1

[68] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 2010. 7, 12

[69] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2021. 1, 3

[70] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *ICCV*, pages 4230–4239, 2023. 7, 8

[71] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021. 1

[72] Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. Pfgm++: Unlocking the potential of physics-inspired generative models. *arXiv preprint arXiv:2302.04265*, 2023. 1

[73] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 1

[74] Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin Cui. Improving diffusion-based image synthesis with context prediction. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[75] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024. 1

[76] Ling Yang, Zhilong Zhang, Zhaochen Yu, Jingwei Liu, Minkai Xu, Stefano Ermon, and Bin CUI. Cross-modal contextualized diffusion models for text-guided visual generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[77] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv*, 2021. 8

[78] Qinsheng Zhang and Yongxin Chen. Diffusion normalizing flow. *Advances in Neural Information Processing Systems*, 34:16280–16291, 2021. 1

[79] Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. *arXiv preprint arXiv:2402.12908*, 2024. 1