

# SCott: Accelerating Diffusion Models with Stochastic Consistency Distillation

Hongjian Liu<sup>1\*</sup>, Qingsong Xie<sup>2\*†</sup>, Tianxiang Ye<sup>3</sup>, Zhijie Deng<sup>3†</sup>, Chen Chen<sup>2</sup>, Shixiang Tang<sup>4</sup>,  
Xueyang Fu<sup>1</sup>, Haonan Lu<sup>2</sup>, Zheng-Jun Zha<sup>1</sup>

<sup>1</sup> University of Science and Technology of China, China

<sup>2</sup> OPPO AI Center

<sup>3</sup> Shanghai Jiao Tong University, China

<sup>4</sup> The Chinese University of Hong Kong

jeffeey@mail.ustc.edu.cn, {xieqingsong1, chenchen4, luhaonan}@oppo.com, {coulson\_tx, zhijied}@sjtu.edu.cn,  
{xyfu, zhazj}@ustc.edu.cn, shixiangtang@cuhk.edu.hk,

## Abstract

The iterative sampling procedure employed by diffusion models (DMs) often leads to significant inference latency. To address this, we propose Stochastic Consistency Distillation (SCott) to enable accelerated text-to-image generation, where high-quality and diverse generations can be achieved within just 2-4 sampling steps. In contrast to vanilla consistency distillation (CD) which distills the ordinary differential equation solvers-based sampling process of a pre-trained teacher model into a student, SCott explores the possibility and validates the efficacy of integrating stochastic differential equation (SDE) solvers into CD to fully unleash the potential of the teacher. SCott is augmented with elaborate strategies to control the noise strength and sampling process of the SDE solver. An adversarial loss is further incorporated to strengthen the consistency constraints in rare sampling steps. Empirically, on the MSCOCO-2017 5K dataset with a Stable Diffusion-V1.5 teacher, SCott achieves an FID of 21.9 with 2 sampling steps, surpassing that of the 1-step InstaFlow (23.4) and the 4-step UFOGen (22.1). Moreover, SCott can yield more diverse samples than other consistency models for high-resolution image generation, with up to 16% improvement in a qualified metric.

**Extended version** — <https://arxiv.org/abs/2403.01505>

## Introduction

Diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song et al. 2020) have emerged as a pivotal component in the realm of generative modeling, facilitating notable progress in domains including image generation (Ramesh et al. 2022; Rombach et al. 2022), video synthesis (Blattmann et al. 2023; Ho et al. 2022), and beyond. In particular, latent diffusion models (LDMs), such as Stable Diffusion (Rombach et al. 2022), have exhibited exceptional capabilities for high-resolution text-to-image synthesis and are acting as fundamental building components for a wide spectrum of downstream applications (Gal et al. 2022; Chen et al. 2023b; Mou et al. 2023).

\*These authors contributed equally. Work done during Hongjian Liu’s internship at OPPO.

†Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

However, it is widely recognized that DMs’ iterative reverse sampling process leads to slow inference. One remediation is improving the solvers used for discretizing the reverse process (Song, Meng, and Ermon 2020; Lu et al. 2022a,b), but it is still hard for them to generate within a limited number of steps (e.g., 5) due to the inevitable discretization errors. Another strategy involves distilling an ordinary differential equation (ODE) based generation process of a pre-trained DM into a shorter one (Salimans and Ho 2022; Meng et al. 2023), but the cost of such progressive distillation is routinely high. Alternatively, consistency distillation (CD) trains unified consistency models (CMs) to fit the consistency mappings characterized by the diffusion ODE for few-step generation (Song et al. 2023). Latent consistency model (LCM) (Luo et al. 2023a) further applies CD to the latent space of a pre-trained autoencoder to enable high-resolution image generation. However, its sample quality is poor within 2 sampling steps. Recently, InstaFlow (Liu et al. 2023), UFOGen (Xu et al. 2024), and ADD (Sauer et al. 2023b) have succeeded in faithfully generating high-resolution images in just 1-2 steps, but they share the limitation of failing to trade additional sampling steps for improved outcomes.

One-step samplers are hard to generate satisfactory outputs, and numerous works (Luo et al. 2023a; Salimans and Ho 2022; Meng et al. 2023) improve their one-step performance via additional steps with acceptable cost (e.g., 4 steps). Our objective is to strike a unified model capable of generating high-quality outputs with 2-4 steps. This is based on the fact that over-emphasizing the one-step generation capacity would unavoidably bias the DM, hence weakening the multi-step generation capacity. Sampling with 2-4 steps does not substantially increase the practical cost compared to 1 step but is likely to improve the upper bound of the sampling quality significantly. We base our solution on CMs because they enjoy the cost-quality trade-off by alternating denoising and noise injection at inference time. Yet, current CD approaches have not fully unleashed the potential of the teacher, considering that for a well-trained DM, ODE-based solvers usually underperform stochastic differential equation (SDE) ones with adequate sampling steps (Xu et al. 2023; Karras et al. 2022; Gonzalez et al. 2023). This is verified by results in Table 1. Empowered by these, we aim to develop Stochastic

**Consistency Distillation (SCott)**, to combine CD with SDE solvers to accelerate the sampling of high-resolution images.

By (Song et al. 2023), CMs are originally defined and learned based on ODE solvers. Naturally, a straightforward adoption of regular SDE solvers suffers from low training stability and poor convergence. To address this, we provide a theoretical justification for using SDE solvers for CD and empirically identify several critical factors that render SCott workable. On one hand, we find it necessary to keep the injected noise in SDE solvers at a moderate intensity to stabilize training while enjoying benefits from the stochasticity inherent in SDE. On the other hand, we find it vital to extend the one-step sampling strategy employed in vanilla CD to a multi-step one to further diminish the discretization errors. The multi-step SDE solver also aids in correcting the accumulated errors in the sampling path, thanks to the injection of random noise (Xu et al. 2023). With these, we obtain a stronger and more versatile teacher for CD.

Surprisingly, we find the uncertainty in SDE can lead to diverse generations for the distilled student model (see Appendix 5). From another perspective, the SDE solver implicitly makes stochastic data augmentation for CD, which aids in mode coverage. To mitigate the imperfect consistency constraints provided by L2 loss in CD, an adversarial learning loss is incorporated to correct student output, further boosting the sample quality at 1-4 sampling steps. Extensive experiments validate the efficacy of SCott in generating high-quality images with conspicuous details. On MSCOCO-2017 5K validation dataset with a Stable Diffusion-V1.5 (SD1.5) (Rombach et al. 2022) teacher, our 2-step method achieves an FID (Heusel et al. 2017) of 21.9, surpassing the previous state-of-the-art, e.g., 2-step LCM (Luo et al. 2023a) (30.4) and 1-step InstaFlow (Liu et al. 2023) (23.4) and UFOGen (Xu et al. 2024) (22.5). Besides, SCott can smoothly improve the sample quality with increasing sampling steps. At a 4-step inference, SCott consistently outperforms LCM, InstaFlow, and UFOGen with 2 – 4 steps in terms of both FID and CLIP Score (Hessel et al. 2021). On MJHQ-5K validation dataset with Realistic-Vision-v51 (RV5.1)<sup>1</sup> teacher, 2-step SCott surpasses LCM by a remarkable margin (24.9 v.s. 37.2 in FID, 0.301 v.s. 0.296 in CLIP Score). For the Coverage metric measuring sample diversity (Naeem et al. 2020), 2-step SCott obtains 0.1232 and 0.1778 gains over 2-step LCM on MSCOCO-2017 5K and MJHQ-5K, respectively.

We summarize our contributions as follows:

- We propose SCott, which accelerates diffusion models to generate high-quality outputs with 1 – 2 steps while maintaining the capability for further improvement via more inference time within 4 steps.
- We provide theoretical convergence analysis for SCott and explore crucial factors that render SCott workable. Furthermore, we integrate adversarial learning objectives into SCott to improve the few-step sample quality.
- SCott achieves 1) a state-of-the-art FID of 21.9 in 2 steps, surpassing competing baselines such as 1-step InstaFlow (23.4), 2-step LCM (30.4), and 4-step UFOGen (22.1),

<sup>1</sup><https://huggingface.co/stablediffusionapi/realistic-vision-v51>.

and 2) much higher sample diversity, reflected by the Coverage metric, than LCM (0.9114 v.s. 0.7882).

## Related Works

**Diffusion Models** Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2020, 2021; Song and Ermon 2020; Karras et al. 2022; Dhariwal and Nichol 2021) progressively perturb data to Gaussian noise and are trained to denoise the noise-corrupted data. During inference, diffusion models create samples from Gaussian distribution by reversing the noising process. They have achieved unprecedented success in text-to-image generation (Saharia et al. 2022; Ramesh et al. 2022), image inpainting (Lugmayr et al. 2022), and image editing (Meng et al. 2021; Chen et al. 2023a). To effectively improve the sample quality of conditioned diffusion models, classifier-free guidance (CFG) (Ho and Salimans 2022) technique is proposed without extra network training.

**Diffusion Acceleration** One of the primary challenges that hinder the practical adoption of diffusion models is the issue of sampling speed due to multiple iterations. Several approaches have been proposed to accelerate the sampling of diffusion models. One type of methods concentrate on training-free numerical solvers (Song, Meng, and Ermon 2020; Lu et al. 2022a,b), such as Denoising Diffusion Implicit Model (DDIM) (Song, Meng, and Ermon 2020) and DPM++ (Lu et al. 2022b). Some researchers explore the approaches of knowledge distillation to compress sampling steps. Progressive Distillation (PD) (Salimans and Ho 2022) and Classifier-aware Distillation (CAD) (Meng et al. 2023) are designed to reduce sampling steps to below 10 steps via multi-stage step distillation. LCM (Luo et al. 2023a) extends CM (Song et al. 2023) to text-to-image generation. However, these methods synthesize blurry samples below four steps. Recently, InstaFlow (Liu et al. 2023), SwiftBrush (Nguyen and Tran 2024), and DMD (Yin et al. 2024) achieve one-step generation in high-resolution text-to-image generation faithfully. InstaFlow proposes a one-step sampling model for text-to-image generation by combining DMs and Rectified Flow (Liu, Gong, and Liu 2022). SwiftBrush adopts variational score distillation (VSD) (Wang et al. 2024) to distill a one-step student. DMD employs distribution matching distillation to enhance the realism of the one-step generator. Nevertheless, they are unable to extend their sampler to multiple steps, and the synthesized images are not satisfactory enough with a single step only. Consistency trajectory models (CTM) (Kim et al. 2023) and Diff-Instruct (Luo et al. 2023b) distill a pre-trained DM into a single-step generator, but their performance on large-scale text-to-image generation is unclear. Motivated by CMs and knowledge distillation, we propose Stochastic Consistency Distillation to generate high-quality images within few steps. Our method is able to not only produce high-quality samples with a 2-step sampler but also improve model performance with increasing steps. Furthermore, due to the introduced stochasticity in SDE solvers, our method exhibits better sample diversity.

**Diffusion GANs** With GANs as a core technique, UFOGen (Xu et al. 2024) proposes a one-step diffusion GAN



Figure 1:  $512 \times 512$  resolution images generated by SCott using 2 sampling steps. SCott is trained based on Realistic-Vision-v51.

for text-to-image generation. However, its image quality can not be significantly improved or even gets worse when increasing steps. Adversarial Diffusion Distillation (Sauer et al. 2023b) distills pre-trained SD models by GANs and score distillation (Poole et al. 2022), achieving one-step generation. However, the training source is not described in the paper, e.g., training time and training data, and hence we can not make a fair comparison with it. Different from UFOGen and ADD which adopt adversarial learning as their core component, we propose to leverage adversarial loss to strengthen consistency constraints in SCott, producing high-quality images at few-step sampling. Consequently, our method inherits the property of CM which heightens text-to-image alignment and image sharpness with increasing sampling steps, and the results is further enhanced by GAN at few-step inference.

### Preliminary

Let  $\mathbf{x} \in \mathbb{R}^k$  denotes a sample from the data distribution  $p_{\text{data}}(\mathbf{x})$  and  $p(\mathbf{z}_t) = \int p_{\text{data}}(\mathbf{x}) \mathcal{N}(\mathbf{z}_t; \alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I}) d\mathbf{x}, \forall t \in [0, T]$  the marginal distribution specified by the forward diffusion process.  $\alpha_t$  and  $\sigma_t$  are positive real-valued functions defining the diffusion schedule so that  $p(\mathbf{z}_0) = p_{\text{data}}(\mathbf{x})$  and  $p(\mathbf{z}_T) \approx \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \tilde{\sigma}^2 \mathbf{I})$  for some  $\tilde{\sigma}$ . A DM  $\epsilon_{\theta}(\cdot, t) : \mathbb{R}^k \rightarrow \mathbb{R}^k$  is trained under score matching principles (Vincent 2011; Song and Ermon 2019; Ho, Jain, and Abbeel 2020) for reversing the diffusion process.<sup>2</sup>

According to the SDE/ODE explanation of the reverse process of DMs (Song et al. 2020), we can obtain an approximate sample of  $p_{\text{data}}(\mathbf{x})$  by drawing a Gaussian noise  $\mathbf{z}_T \sim p(\mathbf{z}_T)$  and then invoking a numerical SDE/ODE solver to discretize the reverse process. Let  $\hat{\mathbf{z}}_t$  denotes the solution at timestep  $t$ , originating from  $\mathbf{z}_T$  based on a solver and model  $\epsilon_{\theta}(\cdot, t)$ , and then  $\hat{\mathbf{z}}_0$  represents the sampled data.

**Consistency Distillation** The mentioned solving process usually hinges on tens or hundreds of steps, causing signifi-

cant inference overheads for practical application. A promising solution is to perform consistency distillation (CD) (Song et al. 2023; Song and Dhariwal 2023) of DMs, yielding a student model  $\mathbf{f}_{\theta}(\cdot, t) : \mathbb{R}^k \rightarrow \mathbb{R}^k$  which enjoys a shortened sampling procedure.

Concretely, CD defines a novel consistency function  $\mathbf{f} : (\mathbf{z}_t, t) \mapsto \mathbf{z}_{\tau}$  for all  $t \in [\tau, T]$ , where  $\tau$  is the boundary time near 0. CD parameterize consistency function with networks  $\mathbf{f}_{\theta}(\mathbf{z}_t, t)$ , and minimizes the following loss for training

$$\min_{\theta} \mathcal{L}_{CD}(\theta) = \mathbb{E}_{n, \mathbf{z}_{t_n}} \left[ \lambda(t_n) \left\| \mathbf{f}_{\theta}(\mathbf{z}_{t_n}, t_n) - \mathbf{f}_{\theta}(\hat{\mathbf{z}}_{t_m}, t_m) \right\|_2^2 \right], \quad (1)$$

where  $t$  in  $[\tau, T]$  are uniformly discretized into  $N$  time points with  $t_1 = \tau < t_2 < \dots < t_N = T$  and  $m \in \{1, \dots, n-1\}$  is a hyper-parameter (Song et al. 2023; Luo et al. 2023a),  $\lambda(\cdot)$  refers to another positive weighting function,  $\theta^-$  denotes the exponential moving average (EMA) of  $\theta$ . The state  $\hat{\mathbf{z}}_{t_m}$  represents an intermediate state, starting from  $\mathbf{z}_{t_n}$  and obtained with the teacher model  $\epsilon_{\theta}(\cdot, t)$  and a one-step DDIM solver (Song, Meng, and Ermon 2020), expressed as

$$\hat{\mathbf{z}}_{t_m} = \frac{\sqrt{\alpha_{t_m}}}{\sqrt{\alpha_{t_n}}} (\mathbf{z}_{t_n} - \sqrt{1 - \alpha_{t_n}} \cdot \epsilon_{\theta}(\mathbf{z}_{t_n}, t_n)) + \sqrt{1 - \alpha_{t_m} - \sigma_{t_n}^2} \cdot \epsilon_{\theta}(\mathbf{z}_{t_n}, t_n) + \sigma_{t_n} \epsilon. \quad (2)$$

CD typically sets  $\sigma_{t_n} = 0$  to keep  $\hat{\mathbf{z}}_{t_m}$  and  $\mathbf{z}_{t_n}$  on the same ODE trajectory.

Although we confine the distance measure in Equation (1) to the squared  $\ell_2$  distance, the  $\ell_1$  distance also applies here. We do not consider the Learned Perceptual Image Patch Similarity (LPIPS, Zhang et al. 2018) because it can lead to inflated FID scores (Song and Dhariwal 2023). The trained  $\mathbf{f}_{\theta}$  allows for one-step or multi-step sampling for generating new data (Song et al. 2023).

**Limitation.** A primary limitation of CD from the viewpoint of distillation is that it has not fully unleashed the potential of the model  $\epsilon_{\theta}(\cdot, t)$ : it utilizes one-step DDIM to sample a preceding state of the current one to serve as the teacher for

<sup>2</sup>The paper focuses on the  $\epsilon$ -prediction type of DMs. Other parameterizations are equivalent in theory (Salimans and Ho 2022).

SOLVER	STEP	FID ↓	CS ↑	CR ↑
DDIM	50	20.3	0.318	0.9130
DPM++	25	20.3	0.318	0.9118
ER-SDE 5	50	<b>20.2</b>	<b>0.320</b>	<b>0.9172</b>

Table 1: Performance comparison of typical ODE solvers (DDIM and DPM++) and SDE ones (ER-SDE 5) on MSCOCO-2017 5K with the SD1.5 model. CS denotes CLIP Score to measure text-to-image consistency and CR represents the Coverage metric to assess sample diversity.

distillation. Yet, for a well-trained DM, ODE-based solvers unusually underperform SDE ones with adequate sampling steps (Gonzalez et al. 2023). Naturally, we ask if we can build an SDE-based teacher for improved CD.

**SDE Solvers for Diffusion Models** The SDE formulation of the reverse-time diffusion process (Song et al. 2020) takes the form of

$$d\mathbf{x}_t = \underbrace{[f_t \mathbf{x}_t + \frac{g_t^2}{2\sigma_t} \epsilon_\theta(\mathbf{x}_t, t)] dt}_{\text{Probabilistic ODE}} + \underbrace{\frac{g_t^2}{2\sigma_t} \epsilon_\theta(\mathbf{x}_t, t) dt + g_t d\bar{\mathbf{w}}_t}_{\text{Langevin process}}, \quad (3)$$

where  $\bar{\mathbf{w}}_t$  denotes the standard Wiener process in reverse time, and

$$f_t := \frac{d \log \alpha_t}{dt}, \quad g_t^2 := \frac{d \sigma_t^2}{dt} - 2 \frac{d \log \alpha_t}{dt} \sigma_t^2. \quad (4)$$

We can discretize Equation (3) over time to get an approximated solution with various SDE solvers (Gonzalez et al. 2023; Cui et al. 2023; Zhang and Chen 2022).

**Benefits of SDE Solvers.** According to the theoretical analysis of (Xu et al. 2023), the divergence between  $p(z_0)$  and the sample distribution  $p(\hat{z}_0)$  stems from the discretization errors along the sampling trajectory and the approximation error between the model and the ground-truth score function (Song et al. 2020). When the number of function evaluations (NFE) is small, SDE solvers exhibit larger discretization errors than ODE ones— $O(\delta^{\frac{3}{2}})$  v.s.  $O(\delta^2)$  with  $\delta$  as the step size for discretization. On the contrary, the discretization errors become less significant as  $\delta$  shrinks and the approximation errors dominate, thus SDE solvers achieve higher sample quality than ODE ones thanks to the injection of noise for correcting previous approximation errors (Karras et al. 2022).

We validate these arguments with results in Table 1, where the ER-SDE 5 solver (Cui et al. 2023) outperforms ODE solvers DDIM and DPM++ (Lu et al. 2022b) in terms of both sample quality and diversity with 50 sampling steps. Besides, the noise injected in SDE solvers implicitly makes stochastic data augmentation for CD. With such observations, we are motivated to explore the possibility of combining SDE solvers and CD, aiming at acquiring stronger teachers for CD.

## Methodology

Previous diffusion acceleration methods either fail to generate high-quality samples within 4 steps (Lu et al. 2022a; Luo

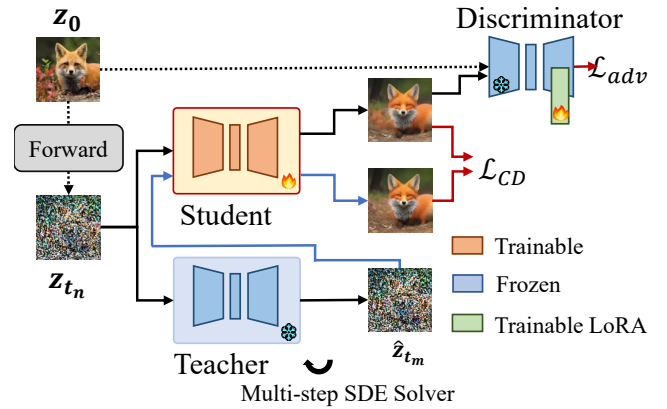


Figure 2: **Overview of SCott.** SCott distills a pre-trained teacher DM into a student one for accelerated sampling. Compared to the vanilla consistency distillation approach, we introduce a multi-step SDE solver to establish a stronger and more versatile teacher. We train the student model with CD loss using SDE solvers. Additionally, we include an adversarial learning loss to correct student output, boosting the sample quality with rare sampling steps. Note that we omit the EMA operation for the teacher for brevity.

et al. 2023a) or are limited to one-step generation, unable to trade additional sampling steps for improved outcomes (Liu et al. 2023; Xu et al. 2024). Our goal is to generate high-quality samples within 2-4 steps to strike a favorable tradeoff between efficiency and efficacy. We observe that CMs enjoy the trade-off by alternating denoising and noise injection at inference time. However, CD approaches haven't unleashed the potential power of the pre-trained teacher DMs, considering ODE solvers underperform SDE ones when the NFE is large. We first lay out a justification for the feasibility of combining CD with SDE solvers. Then, we identify several critical factors to make SCott workable for high-resolution text-to-image generation. We also introduce GAN loss to further enhance the sampling quality at rare steps. Our method is summarized in Figure 2.

## Justification of Using SDE Solvers for CD

Despite our motivation for combining SDE solvers and CD being sensible from the perspective of model distillation, a major concern arises as to whether it is reasonable to do so because CD is originally defined on ODE trajectories. In this section, we provide theoretical justification for this. Following (Song et al. 2023), we provide the convergence proof of using SDE solvers in CD.

**Theorem 1.** Let  $\Delta t := \max_{n \in [1, N]} |t_{n+1} - t_n|$  where  $t \in [\tau, T]$ .

Assume  $f_\theta(\cdot, \cdot)$  is Lipschitz in  $\mathbf{x}$  with constant  $L_1$ . Denote  $f(\cdot, \cdot)$  the consistency function of the SDE defined in Equation (3). Assume the SDE solver  $\Phi_{SDE}$  has a local error bound of  $O((\Delta t)^{p+1})$  with  $p \geq 1$ . Then, if  $\mathcal{L}_{CD}^N(\theta, \Phi_{SDE}) = 0$ , we have :

$$\sup_{n, \mathbf{x}} \|f_\theta(\mathbf{x}_{t_n}, t_n) - f(\mathbf{x}_{t_n}, t_n)\|_2 = O((\Delta t)^p).$$

*Proof.* We refer to Appendix 1 in our supplementary materials for the full proof.  $\square$

Additionally, we also present an empirical investigation to demonstrate the feasibility of using SDE solvers for CD (Please refer to Appendix 3).

### Stochastic Consistency Distillation (SCott)

For large-scale modeling on high-resolution natural images, we, however, observe that directly utilizing an SDE solver for CD leads to training instability and poor convergence. We address this issue by introducing multiple crucial modifications to straightforward implementations.

**Controlling the Level of Noise in SDE Solvers** It is a natural idea to control the level of random noise injected in SDE solvers to stabilize training. Taking the DDIM solver in Equation (2) for example, the  $\sigma_{t_n}$  controls the intensity of injected noise and typically we can introduce a coefficient  $\eta$  for scaling it (Song, Meng, and Ermon 2020):

$$\sigma_{t_n}(\eta) := \eta \sqrt{(1 - \alpha_{t_{n-1}})/(1 - \alpha_{t_n})} \sqrt{1 - \alpha_{t_n}/\alpha_{t_{n-1}}}. \quad (5)$$

Increasing  $\eta$  from 0 to a positive value results in a set of SDE solvers with increasingly intensive noise, starting from an ODE one. We empirically experiment with the solvers for SCott and present the results in Table 5. As shown, we indeed need to select a reasonable noise level to enjoy the benefits of SDE solvers for error correction while avoiding introducing excessive variance. Following this insight, we include the more advanced ER-SDE solver (Cui et al. 2023) to SCott. The novel noise scaling function in the ER-SDE solver also provides a simple way to control the intensity of noise, which aligns with our requirements for SDE solvers.

**Multi-step Sampling** The discretization errors of SDE solvers are larger than those of ODE solvers (Xu et al. 2023). A simple remediation to this is to decrease the step size, but doing so leads to slow convergence and degraded results for SCott, consistent with the results in (Luo et al. 2023a). To address this, we propose a multi-step sampling strategy for SCott. Specifically, for a sample  $z_{t_n}$  at the time  $t_n$  on the SDE trajectory, to obtain the estimated state  $\hat{z}_{t_m}$  for a preceding timestep  $t_m$ , we split the time interval  $t_n - t_m$  into  $h$  intervals, namely, sampling with a step size of  $\frac{t_n - t_m}{h}$ , as illustrated in Figure 3. We denote the solution as  $\hat{z}_{t_m} = \Phi_{SDE}(\epsilon_\theta, z_{t_n}, t_n, t_m, h)$ .

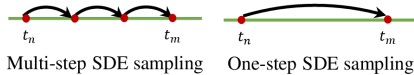


Figure 3: Multi-step SDE solver sampling

Through multi-step sampling, we can reduce the discretization errors in SDE solvers as the interval shrinks. Besides, the multi-step SDE solver aids in correcting the accumulated errors in the sampling path with the injection of random noise. Empirically, we set  $t_m = t_n - 24$  and  $h = 3$ .

**Introducing GAN into SCott** For high-resolution text-to-image generation, considering the high data dimensionality and complex data distribution, simply using L2 loss fails to capture data discrepancy precisely, thus providing imperfect consistency constraints. Suggested by CTM (Kim et al. 2023) which leverages GAN to improve trajectory estimation in the process of distillation, we integrate GAN into SCott to mitigate the restriction of L2 loss. However, it is non-trivial to directly employ CTM’s GAN. Firstly, CTM’s GAN performs on the pixel level while SCott performs on the latent space. For high-resolution image generation, employing a GAN that operates directly on pixel values would significantly increase the training cost. Secondly, SCott is designed to accelerate text-to-image tasks, the interaction between textual information and visual representation is not considered in the design of CTM’s GAN. To address these issues, we design a low-rank adaptation (LoRA) (Hu et al. 2021) discriminator with time and text projection, which renders adversarial training successfully in SCott.

**LoRA Discriminator** To accelerate training, the discriminator is initialized by the pre-trained U-Net in SD1.5. However, we find training the whole U-Net is expensive and unstable. Therefore, we freeze the encoder of the U-Net and only train the LoRA parameters of the decoder of the network. The frozen encoder of the U-Net extracts rich latent representations from the input and conditions, while the LoRA decoder acts as discriminator heads to distinguish between real images and generated images. By employing this approach, only 3.6% of the parameters in the U-Net are updated, which not only contributes to computational efficiency but also increases the overall stability of the adversarial training.

**Conditional Discriminator** Since SCott predicts the clean image  $\hat{z}_0$ , we can directly perform adversarial learning between real image  $z_0$  and generated image  $\hat{z}_0$ . However, we find such a naive design makes training unstable for text-to-image tasks. To surmount this challenge, the timestep  $t_n$  and text  $c$  are propagated into the discriminator as conditions.

Following (Sauer et al. 2023a), we train our discriminator  $D_\phi$  with hinge loss:

$$\begin{aligned} \mathcal{L}_{adv}(\theta, \phi) = & \mathbb{E}_{z_0, c} \left[ \max(0, 1 - D_\phi(z_0, c, 0)) \right] \\ & + \mathbb{E}_{z_{t_n}, c, t_n} \left[ \max(0, 1 + D_\phi(\mathbf{f}_\theta(z_{t_n}, c, \omega, t_n), c, t_n)) \right], \end{aligned} \quad (6)$$

where  $\mathbf{f}_\theta$  denotes the SCott model while  $D_\phi$  represents our proposed LoRA Discriminator. Overall, SCott is trained with the following objective:

$$\mathcal{L}_{SCott}(\theta, \phi) = \mathcal{L}_{CD}(\theta, \theta^-; \Phi_{SDE}) + \lambda_{adv} \mathcal{L}_{adv}(\theta, \phi). \quad (7)$$

In practice, we set  $\lambda_{adv} = 0.4$  to control the strength of the discriminator for refining the outputs of  $\mathbf{f}_\theta$ . To save GPU memory, the training is performed in latent space. We find the introduction of GAN loss renders SCott to produce more realistic outputs, particularly at rare steps. We present more details of the training process for SCott in Appendix 4.

METHOD	REFERENCES	STEP	TIME (s)	FID ↓	CS ↑	CR ↑
DPM++ (LU ET AL. 2022B)	ARXIV (2022)	25	0.88	20.1	0.318	0.9118
DDIM (SONG, MENG, AND ERMON 2020)	ICLR (2021)	50	–	20.3	0.318	0.9130
PD (SALIMANS AND HO 2022)	ICLR (2022)	4	0.21	26.4	0.300	–
CAD (MENG ET AL. 2023)	CVPR (2023)	8	0.34	24.2	0.300	–
LCM (LUO ET AL. 2023A)	ARXIV (2023)	2	0.10	30.4	0.293	0.7882
INSTAFLOW-0.9B (LIU ET AL. 2023)	ICLR (2024)	1	0.09	23.4	0.304	–
INSTAFLOW-1.7B (LIU ET AL. 2023)	ICLR (2024)	1	0.12	22.4	0.309	–
UFOGEN (XU ET AL. 2024)	CVPR (2024)	1	0.09	22.5	0.311	–
UFOGEN (XU ET AL. 2024)	CVPR (2024)	4	–	22.1	0.307	–
SCOTT (OURS)		1	0.09	26.8	0.295	0.8254
SCOTT (OURS)		2	0.12	21.9	0.310	0.9114
SCOTT (OURS)		4	0.18	<b>21.8</b>	<b>0.311</b>	<b>0.9145</b>

Table 2: Comparisons with the state-of-the-art methods on MSCOCO-2017 5K in terms of FID, CS, CR with backbone SD1.5.

METHOD	STEP	FID ↓
SWIFTBRUSH (NGUYEN AND TRAN 2024)	1	16.67
UFOGEN (XU ET AL. 2024)	1	12.78
INSTAFLOW-0.9B (XU ET AL. 2024)	1	13.10
DMD (YIN ET AL. 2024)	1	11.49
SCOTT (OURS)	2	11.13
SCOTT (OURS)	4	<b>10.68</b>

Table 3: Comparisons with the state-of-the-art methods on MSCOCO-2014 30K in terms of FID with backbone SD1.5.

METHOD	STEP	FID ↓	CS ↑	CR ↑
DDIM	50	22.9	0.320	0.8536
LCM	2	37.2	0.296	0.7016
SCOTT (OURS)	2	<b>24.9</b>	<b>0.301</b>	<b>0.8794</b>

Table 4: Comparisons with the state-of-the-art methods on MJHQ-5K in terms of FID, CS, CR, with backbone RV5.1.

## Experiments

In this section, we elaborate on the experimental results of our proposed SCott model for the text-to-image generation task. We start with the comparison with other works, and then ablate components of SCott, highlighting the effectiveness of our proposed components.

**Implementation Details** We use LAION-Aesthetics-6+ dataset (Schuhmann et al. 2022). We train SCott with 4 A100 GPUs and a batch size of 40 for 40K iterations. The learning rate is  $8e-6$  for SCott and  $2e-5$  for the discriminator. (See more experimental setups in Appendix 4)

**Comparison on MSCOCO-2017 5K with SD1.5** To kick-start the comparisons with the state-of-the-art methods, we evaluate the MSCOCO-2017 5K validation dataset (Lin et al. 2014). Zero-shot FID and CLIP Score (CS) with ViT-g/14 backbone (Radford et al. 2021) are exploited as objective metrics. To measure diversity, Coverage (CR) (Naeem et al. 2020) is used as another metric. Table 2 summarizes the performance of our SCott and comparative methods consisting of PD (Salimans and Ho 2022), CAD (Meng et al. 2023), LCM (Luo et al. 2023a), InstaFlow (Liu et al. 2023), UFOGen (Xu et al. 2024). Since PD (Salimans and Ho 2022), CAD (Meng et al. 2023), InstaFlow (Liu et al. 2023), and UFOGen (Xu et al. 2024) do not list CR value, we leave them –. Time denotes inference time on a single A100. LCM is implemented according to the official in our setting.

Among all the methods, our 2-step SCott presents superior FID and CS values than 4-step PD, 8-step CAD, 1-step InstaFlow-0.9B, 2-step LCM. 2-step SCott achieves better

FID than 1-step UFOGen and comparable CS. It is impressive to see that our 2-step SCott beats InstaFlow-1.7B which doubles the parameter size. These results demonstrate our SCott significantly enhances the quality of the generated images while reducing inference steps. The improvements lie in that the proposed  $\mathcal{L}_{CD}$  in SCott is able to reduce the sampling steps remarkably and adversarial loss further improves image quality. Interestingly, the CS decreases by 0.004 when the inference step increases from 1 to 4 for UFOGen. Meanwhile, by changing the inference step from 1 to 4, all the metrics are improved for our SCott, and 4-step SCott outperforms 4-step UFOGen regarding both FID and CS. The result indicates the improvement is limited for UFOGen when increasing steps while our method is much more powerful in enhancing image quality with additional steps in a flexible manner, which is meaningful in the scenarios requiring high-quality images with more affordable computational budgets. The proposed SCott obtains higher CR than LCM, and even outperforms the ODE solvers, 25-step DPM++, and 50-step DDIM, indicating our SCott successfully increases sample diversity due to the introduced randomness in SDE solvers.

We present the qualitative comparisons involving InstaFlow and LCM (See Figure 5 in Appendix 5). Notably, 2-step SCott gains significant improvements over 2-step LCM and 1-step InstaFlow in terms of image quality and text-to-image alignment. We also observe that our generated images exhibit sharper textures and finer details, compared to the images generated by InstaFlow and LCM.

**Comparison on MSCOCO-2014 30K with SD1.5** For complete comparisons, we also benchmark our method against DMD, SwiftBrush, and UFOGen on MSCOCO-2014 30K (Lin et al. 2014). We follow the evaluation setup in

SDE SOLVER	FID ↓	CS ↑	CR ↑
DDIM ( $\eta = 0$ )	27.4	0.296	0.8450
DDIM ( $\eta = 0.1$ )	26.2	0.297	0.8713
DDIM ( $\eta = 0.2$ )	25.2	0.299	0.8786
DDIM ( $\eta = 0.3$ )	27.7	0.298	0.8717
DDIM ( $\eta = 0.6$ )	29.7	0.298	0.8162
ER-SDE 5	<b>24.9</b>	<b>0.301</b>	<b>0.8794</b>

Table 5: Comparisons of different solvers in CD on MJHQ-5K with 2-step inference. All models are based on RV5.1.

DMD and use a CFG scale of 3 following DMD. As depicted in Table 3, our 2-step SCott surpasses all the above methods in FID. Additionally, it’s worth noting that these methods can not improve output with more steps, whereas SCott offers the feasibility to enhance the sample quality with additional steps. With 4-step inference, SCott achieves a superior FID of 10.68. These results validate the effectiveness of our approach in achieving higher-quality samples within 2 steps and improving performance by leveraging additional steps.

**Comparison on MJHQ-5K with RV5.1** To better assess the quality of produced images, we further conduct experiments on MJHQ-5K, randomly selected from MJHQ-30K<sup>3</sup>, since it owns high image quality and image-text alignment, and the correlation between human preference and FID score on the MJHQ is verified by user study. The evaluation of several methods with RV5.1 backbone are listed in Table 4. The reason for distilling RV5.1 is that RV5.1 is much stronger than SD1.5 for text-to-image consistency, which can be found in Appendix 5. Since the training codes of PD, CAD, InstaFlow, and UFOGen are unavailable, they are not included. The result again presents our SCott outperforms LCM by a large margin, because our SDE-based CD provides a stronger and more versatile teacher than ODE-based CD in LCM, and the student output is further refined to be real data by the proposed discriminator.

**Ablation Studies** To analyze the key components of our method, we make a thorough ablation study to verify the effectiveness of the proposed SCott.

**SDE Solver.** Table 5 depicts the results using deterministic and stochastic solvers for CD. DDIM ( $\eta$ ) represents DDIM solver with noise coefficient  $\eta$ , where  $\eta$  denotes the hyperparameter that controls the strength of random noise injected in Eq.2.  $\eta$  achieves an interpolation between the deterministic DDIM ( $\eta = 0$ ) and original DDPM ( $\eta = 1$ ). As observed in Table 5, the SDE-based DDIM ( $\eta = 0.1, 0.2$ ) surpasses ODE-based DDIM ( $\eta = 0$ ) for CD, demonstrating the superiority of SDE solver for CD. This is because the injection of noise at moderate intensity in SDE solvers aids in correcting estimated errors of the teacher model with multiple sampling steps, leading to a more powerful teacher. However, excessive noise intensity leads to poor convergence and degraded samples, as shown in DDIM ( $\eta = 0.6$ ). These results indicate it is crucial to control the noise strength in SCott since large

<sup>3</sup><https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic>.

SOLVER STEP	FID ↓	CS ↑	CR ↑
1	27.4	0.299	0.8461
2	25.4	0.300	0.8724
3	<b>24.9</b>	<b>0.301</b>	<b>0.8794</b>

Table 6: Comparisons of choosing different SDE solver steps during training. All models are based on RV5.1 with 2-step inference on MJHQ-5K.

Loss	FID ↓	CS ↑	CR ↑
SCOTT (Without GAN)	25.0	0.300	0.867
LCM	30.4	0.293	0.788
LCM + GAN	29.2	0.297	0.8160
SCOTT (Full Discriminator)	23.5	0.302	0.8880
SCOTT (Without Condition)	26.1	0.297	0.8412
SCOTT	<b>21.9</b>	<b>0.310</b>	<b>0.9145</b>

Table 7: The performance comparisons of Discriminator on MSCOCO-2017 5k with backbone SD 1.5.

noise leads to low training stability. The 2-order ER-SDE 5 solver further enhances the performance, this is because 1. the 2-order ER-SDE 5 exhibits smaller discretization errors compared to the 1-order DDIM solver. 2. the noise function in ER-SDE 5 mitigates excessive noise, which leads to good convergence. Table 7 shows that SCott without GAN also surpasses the ODE solver-based LCM, further illustrating the benefits of using SDE solvers for CD.

**Multi-step SDE Solver Sampling.** As outlined in Table 6, we study the sampling steps in the process of estimating  $\hat{z}_{t_m}$  given  $z_{t_n}$ . The results indicate that multi-step SDE solvers are superior to single-step solvers. The reason is that SDE solvers usually require multiple iterations to reach the correct destination, demonstrating the choice of SDE solver steps during CD is critical to make SCott successful.

**Discriminator.** Table 7 illustrates our discriminator leads to gains using different inference steps. We observe our LoRA discriminator achieves better sample qualities than the fully parameterized one. Compared with the discriminator without condition, SCott leverages both the time and text conditions. Such a design helps the discriminator better distinguish between real and generated samples. Interestingly, we observe that our discriminator also enhances LCM, which indicates the generalization capability of our discriminator.

## Conclusions

In this paper, we propose stochastic consistency distillation (SCott), a novel approach for accelerating text-to-image diffusion models. SCott integrates SDE solvers into consistency distillation to unleash the potential of the teacher, implemented by controlling noise strength and sampling step of SDE solvers. Adversarial learning is further utilized to aid SCott in generating high-quality images in rare-step sampling. SCott is capable of yielding high-quality images with 2 steps only, surpassing 2-step LCM, 1-step InstaFlow, and 4-step UFOGen. Additionally, SCott consistently improves performance by increased inference cost within 4 steps and exhibits higher diversity than competing baselines.

## References

- Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.
- Chen, J.; Pan, Y.; Yao, T.; and Mei, T. 2023a. Controlstyle: Text-driven stylized image generation using diffusion priors. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7540–7548.
- Chen, W.; Hu, H.; Li, Y.; Rui, N.; Jia, X.; Chang, M.-W.; and Cohen, W. W. 2023b. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*.
- Cui, Q.; Zhang, X.; Lu, Z.; and Liao, Q. 2023. Elucidating the solution space of extended reverse-time SDE for diffusion models. *arXiv:2309.06169*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
- Gonzalez, M.; Fernandez, N.; Tran, T.; Gherbi, E.; Hajri, H.; and Masmoudi, N. 2023. SEEDS: Exponential SDE Solvers for Fast High-Quality Sampling from Diffusion Models. *arXiv:2305.14267*.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577.
- Kim, D.; Lai, C.-H.; Liao, W.-H.; Murata, N.; Takida, Y.; Uesaka, T.; He, Y.; Mitsufuji, Y.; and Ermon, S. 2023. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.
- Liu, X.; Zhang, X.; Ma, J.; Peng, J.; and Liu, Q. 2023. InstafLOW: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*.
- Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.
- Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.
- Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.
- Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023a. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.
- Luo, W.; Hu, T.; Zhang, S.; Sun, J.; Li, Z.; and Zhang, Z. 2023b. Diff-Instruct: A Universal Approach for Transferring Knowledge From Pre-trained Diffusion Models. *arXiv preprint arXiv:2305.18455*.
- Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.
- Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14297–14306.
- Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.
- Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable fidelity and diversity metrics for generative models. In *37th International Conference on Machine Learning, ICML 2020*, 7133–7142. International Machine Learning Society (IMLS).
- Nguyen, T. H.; and Tran, A. 2024. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7807–7816.
- Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.
- Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*.
- Sauer, A.; Karras, T.; Laine, S.; Geiger, A.; and Aila, T. 2023a. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning*.
- Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2023b. Adversarial Diffusion Distillation. *arXiv preprint arXiv:2311.17042*.
- Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Song, Y.; and Dhariwal, P. 2023. Improved Techniques for Training Consistency Models. *arXiv preprint arXiv:2310.14189*.
- Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. *arXiv preprint arXiv:2303.01469*.
- Song, Y.; Durkan, C.; Murray, I.; and Ermon, S. 2021. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428.
- Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.
- Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Vincent, P. 2011. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7): 1661–1674.
- Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.
- Xu, Y.; Deng, M.; Cheng, X.; Tian, Y.; Liu, Z.; and Jaakkola, T. 2023. Restart Sampling for Improving Generative Processes. *arXiv preprint arXiv:2306.14878*.
- Xu, Y.; Zhao, Y.; Xiao, Z.; and Hou, T. 2024. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8196–8206.
- Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6613–6623.
- Zhang, Q.; and Chen, Y. 2022. Fast Sampling of Diffusion Models with Exponential Integrator. In *The Eleventh International Conference on Learning Representations*.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.