PROGRESSIVE DATA-FREE DIFFUSION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While one-step distillation achieves strong single-step generation, these methods are not inherently flexible for multi-step sampling. Efforts to adapt them beyond one step frequently lead to a reliance on training data, poor generation quality at early intermediate steps, and significant computational demands. To overcome these limitations, we propose Progressive Multi-step Diffusion Distillation (PMDD), a unified framework that generalizes one-step distillation to the multi-step setting. PMDD adopts a recursive training strategy in which an N-step student is progressively refined into an N+1-step student with minimal finetuning. This process is enabled by a data-free sampling mechanism for generating intermediate states and an unforget loss that maintains the generation quality across steps. Together, these innovations allow PMDD to match or surpass a teacher model with only a handful of function evaluations, while providing scalable, data-free training and substantially reduced computational overhead. Extensive experiments demonstrate that our method not only outperforms established few-step diffusion approaches but also gains teacher-level-exceeded performance, with FID 1.95 on ImageNet 64×64 and FID 8.26 on zero-shot COCO 512×512, making a new state of the art in multi-step data-free distillation with significantly lower resource demands.

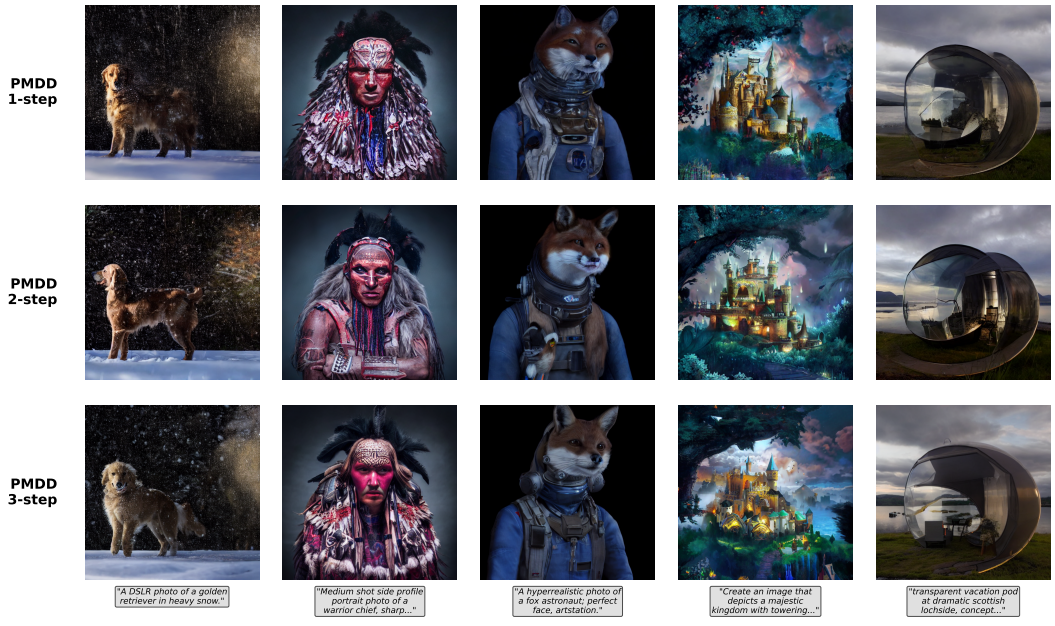

Figure 1: 512×512 samples produced by our 3-step generator distillation of SD v1.5. All images are produced from a single unified model.

1 INTRODUCTION

Diffusion models have achieved remarkable success in generative tasks, demonstrating state-of-the-art performance across a wide range of domains such as image generation (Song et al., 2019; 2020b; Ho et al., 2020; Song et al., 2020a), audio synthesis (Chen et al., 2021; Kong et al., 2021), and text generation (Austin et al., 2021; Gulrajani & Hashimoto, 2023; Lou et al., 2024). Diffusion-based image generation models adopt an iterative denoising process which gradually removes noise from a noisy intermediate sample to reconstruct a high-quality image. However, the sampling process is inherently slow, typically requiring hundreds of neural function evaluations (NFEs), which makes the models expensive for many real-world applications.

To overcome this limitation, recent research has applied the distillation approach to distill a (diffusion) teacher model into a (diffusion) student model. A common strategy is to directly match the deterministic outputs of the teacher's iterative denoising process with those of the student in one or a few steps (Luhman & Luhman, 2021; Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023a; Dao et al., 2024; Kim et al., 2024); though such trajectory-matching still underperforms the teacher. In contrast, distributional-matching methods, motivated by frameworks such as GMNNs (Li et al., 2015) and GANs, bypass trajectory approximation by learning a one-step mapping from noise to clean data, ensuring that the student matches the teacher's overall output distribution (Luo et al., 2024; Nguyen & Tran, 2024; Yin et al., 2024b;a; Zhou et al., 2024). While promising, these methods have several limitations. First, they often lack the flexibility to support multi-step sampling for higher fidelity, which is critical in high-fidelity text-to-image generation where one-step is often insufficient, and multiple steps are required to refine outputs. Second, even when multi-step extensions are possible, they remain strongly dependent on data. Additionally, multistep models when sampled with only a few steps perform poorly, while still requiring substantial computational resources during the training process. ***For example, Multistep Moment Matching*** (Salimans et al., 2024) ***required 256 TPUv5e chips for two weeks of training, DMD v2 consumed 64 A100 GPUs in over a day.***

In this paper, we unify prior one-step diffusion distillation approaches under a general multi-step framework and directly address these bottlenecks. Specifically, to extend the framework beyond a single step, we introduce a progressive training strategy that incrementally expands an N-step teacher model into an $N+1$-step student model, achieving improved generation fidelity with minimal finetuning overhead. However, naively applying this framework introduces two key challenges: (1) maintaining high-quality generation of intermediate latent samples, and (2) avoiding catastrophic forgetting of earlier iterations. To address these challenges, we propose a novel data-free sampling approach for intermediate states, enabling strong performance without requiring external data. To mitigate catastrophic forgetting, we introduce an unforget loss, which preserves the generation quality across iterations and substantially improves the few-step sampling setting.

We evaluate our approach across various tasks, including conditional image generation on CIFAR-10 (Krizhevsky, 2009), ImageNet 64×64 (Russakovsky et al., 2015), and zero-shot text-to-image generation on MS COCO 512×512 (Lin et al., 2014). As shown in experimental results, our one-step model consistently surpasses prior distillation methods, including Diff-Instruct (Luo et al., 2024), Distribution Matching Distillation (Yin et al., 2024b) and Consistency Models (Song et al., 2023), and even teacher models in some cases. In the multi-step setting, PMDD scales predictably with the number of steps, and outperforms other diffusion distillation methods, especially Few-step Score Identity Distillation (Zhou et al., 2025), achieving *a new state-of-the-art* in multi-step data-free distillation with **FID 8.26** on MS-COCO 2014-30k. These results are obtained with far fewer finetuning steps and substantially less computation; PMDD is trained in 5-6 days using at most **3 H100 GPUs**. In addition, compared to teacher models requiring tens or hundreds of NFEs, PMDD achieves comparable FID with 10×-20× higher efficiency.

2 PRELIMINARY

One-step diffusion distillation involves learning a generator $g_\phi(x_T, T)$ (typically referred to as student) that can generate data samples x_0 from Gaussian noise samples $x_T \sim \mathcal{N}(0, \mathrm{I})$ by leveraging a pretrained diffusion model (typically referred to as teacher). A standard approach to this problem is to match the data distributions $p_\phi(x_0)$ and $p_\theta(x_0)$ characterized by g_ϕ and the pretrained teacher,

respectively by minimizing the following KL divergence:

$$\mathcal{L}_{\text{KL}}(\phi) := D_{\text{KL}}(p_\phi(x_0) \| p_\theta(x_0)) \tag{1}$$

However, directly minimizing this KL divergence is difficult. Therefore, in practice, we minimize its variational upper bound (Ho et al., 2020; Song et al., 2020b):

$$\mathcal{L}_{\text{VUB}}(\phi) := \sum_{t=1}^{T} D_{\text{KL}}(p_\phi(x_{t-1}|x_t) \| p_\theta(x_{t-1}|x_t)) \tag{2}$$

Here, $p_\theta(x_{t-1}|x_t)$ is the parameterized backward transition distribution of the teacher while $p_\phi(x_{t-1}|x_t)$ can be regarded as the backward transition distribution of an "imaginary" diffusion model that captures $p_\phi(x_0)$. $p_\phi(x_{t-1}|x_t)$ can be parameterized in the same way as $p_\theta(x_{t-1}|x_t)$ with parameters that can be adapted from those of $p_\theta(x_{t-1}|x_t)$.

Since $p_\theta(x_{t-1}|x_t)$ is typically parameterized as $p(x_{t-1}|x_t, x_\theta(x_t, t))$ where $x_\theta(x_t, t)$ is a parametric approximation of $\mathbb{E}_{p_\theta(x_0|x_t)}[x_0]$ (Song et al., 2020a; Kingma et al., 2021), we can also express $p_\phi(x_{t-1}|x_t)$ as $p(x_{t-1}|x_t, x_\phi(x_t, t))$ with $x_\phi(x_t, t)$ approximating $\mathbb{E}_{p_\phi(x_0|x_t)}[x_0]$. Consequently, minimizing $\mathcal{L}_{\text{VUB}}(\phi)$ becomes minimizing the denoised-sample matching (DM) loss $\mathcal{L}_{\text{DM}}(\phi)$ below:

$$\mathcal{L}_{\text{DM}}(\phi) := \mathbb{E}_{x_0^T \sim g_\phi(x_T, T), t, \epsilon, x_t}\left[w_x(t) \|x_\phi(x_t, t) - x_\theta(x_t, t)\|_2^2\right] \tag{3}$$

where $x_T \sim \mathcal{N}(0, \mathrm{I})$, $t \sim \mathcal{U}(1, T)$, $\epsilon \sim \mathcal{N}(0, \mathrm{I})$, $x_t = a_t x_0^T + \sigma_t \epsilon$, and $w_x(t) > 0$ denotes the time-dependent loss coefficient w.r.t. the denoised-sample parameterization.

The main challenge when minimizing this loss is that $x_\phi(x_t, t)$ is unknown. One way to get around this problem is replacing it with the following surrogate loss (Poole et al., 2022):

$$\tilde{\mathcal{L}}_{\text{NM}}(\phi) := \mathbb{E}_{x_0^T \sim g_\phi(x_T, T), t, \epsilon, x_t}\left[w_\epsilon(t)(\epsilon_\theta(x_t, t) - \epsilon)\right]$$

where $\epsilon_\theta(x_t, t)$ can be derived from $x_\theta(x_t, t)$ and x_t via Tweedie's formula (Efron, 2011). However, minimizing $\tilde{\mathcal{L}}_{\text{NM}}(\phi)$ is not equivalent to minimizing the KL divergence between $p_\phi(x_t)$ and $p_\theta(x_t)$ in Eq. 1. Consequently, this loss can lead to low-quality and low-diversity samples from the student network g_ϕ, as observed in (Wang et al., 2024).

A better approach is to find a good approximation of $x_\phi(x_t, t)$ in Eq. 3. This can be done by training an adapted denoising network x_φ on clean samples generated by g_ϕ, using the following loss:

$$\mathcal{L}_{\text{adapted}}(\varphi) = \mathcal{L}_{\text{DM}}(\varphi) := \mathbb{E}_{x_0^T \sim g_\phi(x_T, T), t, \epsilon, x_t}\left[w_x(t) \left\|x_\varphi(x_t, t) - x_0^T\right\|_2^2\right] \tag{4}$$

After training x_φ, we update g_ϕ using a version of $\mathcal{L}_{\text{DM}}(\phi)$ with $x_\phi(x_t, t)$ replaced by $x_\varphi(x_t, t)$:

$$\mathcal{L}_{\text{DM}}(\phi) \approx \mathbb{E}_{x_0^T \sim g_\phi(x_T, T), t, \epsilon, x_t}\left[w_x(t) \|x_\varphi(x_t, t) - x_\theta(x_t, t)\|_2^2\right] \tag{5}$$

To stabilize training, prior works (Wang et al., 2024; Nguyen & Tran, 2024; Yin et al., 2024b) replace the full gradient $\nabla_\phi \mathcal{L}_{\text{NM}}(\phi)$ with a modified gradient:

$$\tilde{\nabla}_\phi \mathcal{L}_{\text{DM}}(\phi) := \mathbb{E}_{x_0^T \sim g_\phi(x_T, T), t, \epsilon, x_t}\left[w_x(t)(x_\varphi(x_t, t) - x_\theta(x_t, t))\frac{\partial g_\phi(x_T)}{\partial \phi}\right] \tag{6}$$

In practice, ϵ_φ and g_ϕ are optimized alternately by minimizing $\mathcal{L}_{\text{DM}}(\varphi)$ and updating ϕ with $\tilde{\nabla}_\phi \mathcal{L}_{\text{DM}}(\phi)$. More recently, Zhou et al. (Zhou et al., 2024) introduce the score identity distillation (SiD) loss into $\mathcal{L}_{\text{DM}}(\phi)$, which enables robust and stable training without the need for gradient modification. Their student loss takes the form:

$$\begin{aligned}\mathcal{L}_{\text{student}}(\phi) = &\mathbb{E}_{x_0^T \sim g_\phi(x_T, T), t, \epsilon, x_t}\left[w_x(t) \|x_\varphi(x_t, t) - x_\theta(x_t, t)\|_2^2\right] \\ &+ \alpha \mathbb{E}_{x_0^T = g_\phi(x_T, T), t, \epsilon, x_t}\left[w_x(t)(x_\theta(x_t, t) - x_\varphi(x_t, t))^\top (x_\theta(x_t, t) - x_0^T)\right] \\ &= \mathcal{L}_{\text{DM}}(\phi) + \alpha \mathcal{L}_{\text{SiD}}(\phi)\end{aligned}$$

3 METHOD

Most diffusion distillation methods either rely on the teacher model's original training data Song et al. (2023); Xie et al. (2024); Yin et al. (2024b) or are restricted to one-step distillation Gu et al. (2023); Nguyen & Tran (2024). In contrast, we study a more general and challenging setting: *data-free multistep distillation*. Our goal is to train an n-step student model g_ϕ capable of generating clean samples x_0 from any time steps t_i ($1 \le i \le n$) under the constraint $0 < t_1 < t_2 < \ldots < t_n = T$, all without any access to clean training data.

The key difficulty lies in obtaining intermediate samples $x_{t_i} \sim p(x_{t_i})$ for $t_i < T$. With clean data, this is trivial: draw x_0 from the dataset and then generate x_{t_i} via the forward process $p(x_{t_i}|x_0)$. In the one-step case, we can directly sample from $\mathcal{N}(0, \mathrm{I})$. Unfortunately, neither of these options applies in the data-free multistep scenario.

A naive solution is to sample $x_T \sim \mathcal{N}(0, \mathrm{I})$ and run the teacher's backward process to obtain x_{t_i}. Yet this simulation-based approach becomes computationally expensive as t_i approaches 0, making large-scale training impractical. To overcome this, we propose a *progressive distillation* strategy, where the student g_ϕ is distilled in multiple stages, sequentially from t_n down to t_1. Concretely, in the first stage ($t_n = T$), we train g_ϕ with Gaussian inputs using the distillation framework described in Section 2. Once g_ϕ can generate clean samples from step t_n down to t_{i+1}, we further adapt it to handle step t_i, repeating this process until reaching t_1. To sample x_{t_i}, we begin with $x_{t_n} \sim \mathcal{N}(0, \mathrm{I})$ and recursively apply:

$$x_0^{t_k} = \mathrm{sg}\left(g_\phi\left(x_{t_k}, t_k\right)\right), \;\; x_{t_{k-1}} = a_{t_{k-1}} x_0^{t_k} + \sigma_{t_{k-1}} \epsilon \tag{7}$$

where k runs from n to $i+1$, $\epsilon \sim \mathcal{N}(0, \mathrm{I})$, and sg denotes the stop-gradient operator. Since each $x_0^{t_k}$ approximates samples from $p(x_0)$, the resulting x_{t_i} closely follows $p(x_{t_i})$. We then pass x_{t_i} through g_ϕ(with gradients enabled) to obtain $x_0^{t_i}$ and alternately optimize g_ϕ and the adapted denoising network x_φ under the following distillation objectives:

$$\mathcal{L}^i_{\text{adapted}}(\varphi) = \mathbb{E}_{x_0^{t_i} = g_\phi\left(x_{t_i}, t_i\right), t, \epsilon, x_t}\left[w_x(t)\left\|x_\varphi(x_t, t) - x_0^{t_i}\right\|_2^2\right] = \mathcal{L}^i_{\text{DM}}(\varphi) \tag{8}$$

$$\begin{aligned}
\mathcal{L}^i_{\text{student}}(\phi) = {} & \mathbb{E}_{x_0^{t_i} = g_\phi\left(x_{t_i}, t_i\right), t, \epsilon, x_t}\left[w_\epsilon(t)\left\|x_\varphi(x_t, t) - x_\theta(x_t, t)\right\|_2^2\right] \\
& + \alpha \mathbb{E}_{x_0^{t_i} = g_\phi\left(x_{t_i}, t_i\right), t, \epsilon, x_t}\left[w_x(t)\left(x_\theta(x_t, t) - x_\varphi(x_t, t)\right)^\top\left(x_\theta(x_t, t) - x_0^{t_i}\right)\right] \\
& + \beta \sum_{k=i+1}^{n} \mathbb{E}_{x_{t_k}}\left[\left\|g_\phi\left(x_{t_k}, t_k\right) - g^{t_k}_{\text{old}}\left(x_{t_k}, t_k\right)\right\|_2^2\right] \qquad (9) \\
= {} & \mathcal{L}^i_{\text{DM}}(\phi) + \alpha \mathcal{L}^i_{\text{SID}}(\phi) + \beta \mathcal{L}^i_{\text{unforget}}(\phi) \qquad (10)
\end{aligned}$$

Here, $t \sim \mathcal{U}(1, T)$, $\epsilon \sim \mathcal{N}(0, \mathrm{I})$, and $x_t = a_t x_0^{t_i} + \sigma_t \epsilon$. The last term in Eq. 9 plays a critical role in preventing g_ϕ from catastrophically forgetting the learned multi-step mappings. In this term, $g^{t_k}_{\text{old}}$ is the previous version of the student distilled at step t_k up to step t_{i+1}.

4 EXPERIMENT

We assess the effectiveness of our method for distilling pretrained diffusion models on both class-conditional image generation and text-to-image generation tasks. For class-conditional generation, we adopt CIFAR-10 (Krizhevsky, 2009) and ImageNet 64×64 (Russakovsky et al., 2015) as benchmarks, using the pretrained teacher models from (Karras et al., 2022). For text-to-image generation, we distill from a pretrained Stable Diffusion v1.5 (Rombach et al., 2022) and evaluate on MS-COCO 30k (Lin et al., 2014), following standard practice in prior work (Yin et al., 2024b; Salimans et al., 2024; Zhou et al., 2025).

4.1 CLASS-CONDITIONAL IMAGE GENERATION

We benchmark PMDD against recent diffusion distillation methods on CIFAR-10 32×32 and ImageNet 64×64. We follow the implementation of DMD (Yin et al., 2024b), where we generate

Type	Method	NFE (↓)	FID (↓)
Teacher	VP-EDM (Karras et al., 2022)	35	1.79
One-step	GET-Base (Yin et al., 2024b)	1	6.25
	Meng et al. (Meng et al., 2023)	1	5.98
	DMD (w/o reg.) ✓ (Yin et al., 2024b)	1	5.58
	Diff-Instruct ✓ (Luo et al., 2024)	1	4.19
	DMD (w/o KL) (Yin et al., 2024b)	1	3.82
	DMD (Yin et al., 2024b)	1	2.66
	SiD ($\alpha = 1.0$) ✓ (Zhou et al., 2024)	1	1.93
Multi-step	Progressive Distillation† (Salimans & Ho, 2022)	1	9.12
		2	4.51
	Consistency Distillation† (Song et al., 2023)	1	3.55
		2	2.93
	CTM (Kim et al., 2024)	1	1.73
		2	1.63
	PMDD (ours) ✓	1	2.52
		2	2.19

(a) CIFAR-10

Type	Method	NFE (↓)	FID (↓)
Teacher	VP-EDM (Karras et al., 2022)	79	2.64
One-step	BOOT ✓ (Gu et al., 2023)	1	16.3
	DFNO (Zheng et al., 2023a)	1	7.83
	TRACT (Berthelot et al., 2023)	1	7.43
	SwiftBrush ✓ (Nguyen & Tran, 2024)	1	5.85
	Diff-Instruct ✓ (Luo et al., 2024)	1	5.57
	DMD (Yin et al., 2024b)	1	2.62
	DMD v2 (w/o GAN) ✓ (Yin et al., 2024a)	1	2.61
	DMD v2 (Yin et al., 2024a)	1	1.28
	SiD ($\alpha = 1.0$) ✓ (Zhou et al., 2024)	1	2.02
Multi-step	Progressive Distillation (Salimans & Ho, 2022)	1	15.39
		2	8.95
	Consistency Distillation (Song et al., 2023)	1	6.20
		2	4.70
	Moment Matching (Salimans et al., 2024)	1	3.0
		2	3.86
	CTM (Kim et al., 2024)	1	1.92
		2	1.73
	PMDD (ours) ✓	1	2.60
		2	1.95

(b) ImageNet 64×64

Table 1: Results on CIFAR-10 (left) and ImageNet 64×64 (right) of our method and baselines. Data-free distillation methods are marked with ✓, unconditional methods are marked with †.

50,000 images for every 1000 training iterations in order to calculate the FID metric (Heusel et al., 2017), and report the best model achieving the lowest FID during evaluation. At each stage, the student and adapted networks are reinitialized from their best-performing checkpoints. We summarize the results in Table 1.

One-step PMDD achieves an NFE-1 FID of 2.52 on CIFAR-10 and 2.60 on ImageNet 64×64, outperforming prior state-of-the-art data-dependent one-step distillation methods such as TRACT, DFNO, and DMD, as well as recent data-free approaches including SwiftBrush and Diff-Instruct. PMDD ranks only behind Score Identity Distillation (SiD) and CTM; however, CTM is data-dependent, while SiD is substantially more computationally demanding (see discussion below). Compared to pretrained teacher diffusion models such as DDIM, PMDD achieves $\approx 3.3\times$ lower FID while being $10\times$ faster.

Multistep PMDD surpasses established data-dependent multi-step methods, including Progressive Distillation, Consistency Distillation and Multistep Moment Matching, achieving an FID of 1.95 with only two function evaluations (NFE = 2). It also outperforms SiD (FID = 2.02); however, SiD requires the equivalent of 1 billion synthetic training images (121k iterations with a very large batch size), whereas our method achieves competitive performance using only around 34M synthetic images in total - nearly 30× fewer. Furthermore, compared to its pretrained teacher model, PMDD observed $\approx 1.25\times$ lower FID while being $\approx 40\times$ faster and more efficient.

4.2 TEXT-TO-IMAGE GENERATION

To assess the scalability of our approach to large-scale dataset, we distill a latent-space model at 512×512 resolution using Stable Diffusion v1.5 (Rombach et al., 2022) following prior work settings (Yin et al., 2024a). Evaluation is conducted on zero-shot MS COCO, where we report both FID and CLIP score to measure fidelity and text-image alignment.

Table 2 shows that our 3-step PMDD surpasses nearly all diffusion distillation methods, with the only exception of Moment Matching — a data-dependent approach that requires up to 8 NFEs for sampling and massive compute (256 TPUv5 cores for over two weeks of training). In contrast, PMDD achieves an FID of 8.50 with only 3 sampling steps, a data-free method trained in 8 days on 3 H100 GPUs. Remarkably, PMDD even outperforms its teacher model (SDv1.5, 50 NFEs, 8.52 FID). It also outperforms Few-step Score Identity Distillation, a concurrent data-free distillation method, while requiring fewer sampling steps and yielding better FID. These results establish PMDD as the new state of the art in few-step data-free distillation.

Behavior of PMDD under varying inference budgets Table 2 reports the best model trained for each step count. We further analyze the robustness of PMDD when generating images with varying inference budgets under a single unified model. Table 3 shows that in both 2-step and 3-step settings, the unforget loss $\mathcal{L}^i_{\text{unforget}}(\phi)$ yields stronger final-step performance than competing methods, and

Method	NFE (↓)	COCO FID$_{30k}$ (↓)	CLIP Score (↑)
Base Models			
SD v1.5 (CFG = 3) (Rombach et al., 2022)	512	8.78	-
SD v1.5 (CFG = 8) (Rombach et al., 2022)	512	13.45	0.322
Diffusion Distillation (One-step)			
DMD (CFG=3) (Yin et al., 2024b)	1	11.49	-
DMD (CFG=8) (Yin et al., 2024b)	1		0.32
SwiftBrush (Nguyen & Tran, 2024)	1	16.67	0.29
SwiftBrush+PG+NASA (Nguyen et al., 2024)	1	9.94	0.31
InstaFlow-1.7B (Liu et al., 2023)	1	11.8	0.309
DMDv2 (CFG = 1.75) (Yin et al., 2024a)	1	8.35	0.30
Diffusion Distillation (Multistep)			
LCM-LoRA (Luo et al., 2023b)	4	23.62	
PeRFlow (Yan et al., 2024)	4	18.59	-
SLAM (Xu et al., 2024)	4	10.06	
Moment Matching (CFG = 0) (Salimans et al., 2024)	8	7.25	
DMDv2 w/o GAN (CFG = 1.75) (✓) (Yin et al., 2024a)	1	9.35	0.304
(reimplemented)	2	10.44	0.301
	3	9.18	0.303
Few-step Score Identity Distillation (Zero-CFG) (✓) (Zhou et al., 2025)	1	9.63	0.321
	2	8.75	0.315
	4	8.52	0.308
PMDD (CFG = 1.75) (✓)	1	10.41	0.302
PMDD (CFG = 1.75) (✓)	2	8.63	0.30
PMDD (CFG = 1.75) (✓)	3	8.50	0.302
PMDD (CFG = 1.0) (✓)	3	**8.26**	0.298

Table 2: Comparison of image generation methods on 30k COCO-2014 prompts, following a standard evaluation protocol. Methods that are data-free (✓)

Method	NFE=3	NFE=2	NFE=1
Guided Distill.	-	33.25	108.21
LCM	-	13.31	35.36
Self-corrected Flow Distillation	-	11.46	11.91
DMDv2 w/o GAN (reimplemented)	-	10.44	16.22
PMDD (CFG = 1.75)	-	**8.63**	**11.67**
DMDv2 w/o GAN (reimplemented)	9.18	10.36	23.32
PMDD (CFG = 1.75)	**8.50**	10.07	**12.65**

Table 3: FID comparison of diffusion distillation methods under varying sampling budgets

Unforget Weight ($\beta = 1.0$)	External Sampling of x_{t_i}	$\mathcal{L}_{\text{SiD}}$	CIFAR-10	ImageNet 64×64
			5.89	8.01
✓			3.02	3.71
✓	✓		2.94	3.58
✓	✓	✓	2.21	1.99

Table 4: Ablation Study of 2-step model on CIFAR-10 and ImageNet 64×64. FID is reported for all experiments.

maintains high fidelity even at low step counts. We examine the role of $\mathcal{L}^i_{\text{unforget}}(\phi)$ more closely in Section 4.3.

Figure 1 demonstrates that, conditioned on the same initial noise x_T, PMDD consistently preserves a coherent global image structure across different sampling steps. Subsequent steps typically refine fine details, such as facial expressions, while the overall structure remains intact. This shows the possibility of utilizing a single model across all steps, suitable for varying inference budgets depending on available resources and desired generation quality.

4.3 Ablation Studies

We conduct extensive ablation studies on our distilled model, which explores the impact of three key factors: the unforget weight ($\beta = 1.0$), the inclusion of additional score identity loss $\mathcal{L}^i_{\text{SiD}}(\phi)$, and the role of external sampling of x_{t_i} during training and sampling using previously trained models. Table 4 demonstrates that model performance is mainly driven by two key components: the score identity loss $\mathcal{L}^i_{\text{SiD}}(\phi)$ and the unforget loss $\mathcal{L}^i_{\text{unforget}}(\phi)$.

Table 5 indicates that under $\mathcal{L}^i_{\text{DM}}(\phi)$, PMDD's performance shows consistent improvements as the number of sampling steps increases. In contrast, while $\mathcal{L}^{(t_i)}_{\text{SiD}}(\phi)$ yields strong results under 2-step inference, it does not scale effectively to additional steps, limiting further improvements in image quality. Moreover, when applied to higher-dimensional image generation tasks such as Stable Diffusion, $\mathcal{L}^i_{\text{SiD}}(\phi)$ leads to poor performance and fails to learn successfully. Extending $\mathcal{L}^{(t_i)}_{\text{SiD}}(\phi)$ to large-scale text-to-image generation task for PMDD is left for future work.

	CIFAR-10		ImageNet 64×64	
Inference Steps	$\mathcal{L}^i_{\text{DM}}(\phi)$	$\mathcal{L}(\phi) + \alpha\mathcal{L}^i_{\text{SiD}}(\phi)$	$\mathcal{L}^i_{\text{DM}}(\phi)$	$\mathcal{L}(\phi) + \alpha\mathcal{L}^i_{\text{SiD}}(\phi)$
FID (NFE = 1)	3.49	2.52	3.70	2.60
FID (NFE = 2)	2.91	2.19	3.58	1.95
FID (NFE = 3)	2.86	2.48	3.47	2.14

Table 5: Ablation of the loss term on distilling a 2-step model on CIFAR-10 and ImageNet 64×64. By default, we use our best hyper-parameters $\alpha = 1.0$ and $\beta = 0.3$.

Figure 2 further explores the impact of varying the unforget loss weight β on CIFAR-10 and ImageNet 64×64. The effect is minimal for 2-step sampling but becomes significant in learning to unforget 1-step. For Stable Diffusion, Table 6 and Figure 3 indicate that performance is highly sensitive to this weight, with optimal results achieved when $\alpha \in [0.01, 0.1]$, highlighting the critical role of precise loss balancing in our framework. Larger weights overemphasize the unforget objective at the expense of distribution matching loss, preserving fidelity in earlier steps while degrading final-step quality.

Table 6 compares the effect of external sampling of x_{t_i}. In 2-step sampling, the difference between using and not using external sampling is marginal; however, in 3-step sampling the effect is substantial (Figure 3). With CFG = 1.75, training without external sampling (brown line) requires roughly twice as many iterations to match the convergence speed of training with external sampling (red line). This occurs because, without external sampling, the model must learn to map from a constantly changing x_{t_i} (generated by the current model and therefore not fixed), while also handling unforget at earlier steps. In contrast, external sampling fixes x_{t_i}, allowing the model to focus on reducing the FID of the final step, while requiring a larger unforget weight to preserve fidelity at earlier steps.

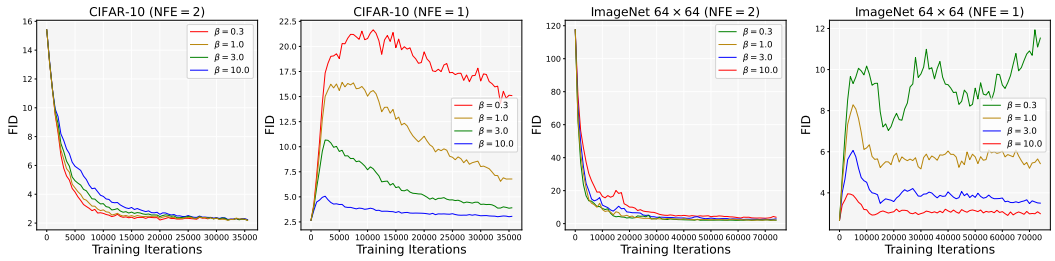

Figure 2: Effect of unforget loss β on 1-step while training 2-step for CIFAR-10 and ImageNet 64×64 ($\alpha = 1.0$)

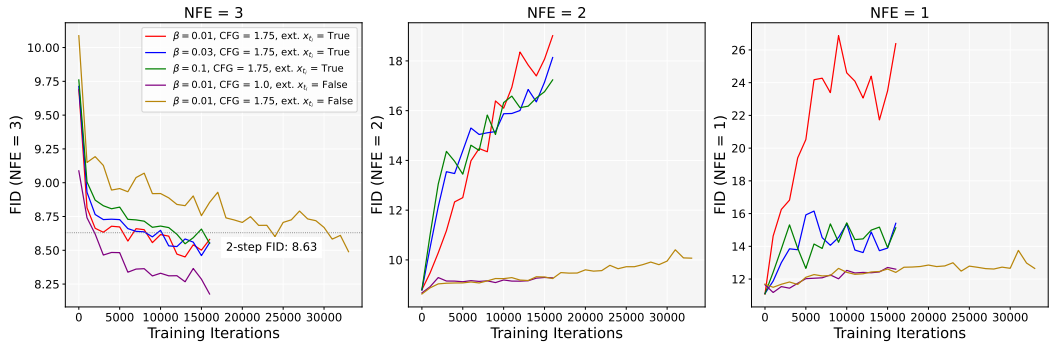

Figure 3: Effect of unforget loss β on 3-step inference for COCO 2014

5 Related Work

Training-free methods employ higher-order numerical solvers to expedite the backward process, especially high-order SDE Solvers. For instance, Stochastic Explicit Exponential Derivative-free

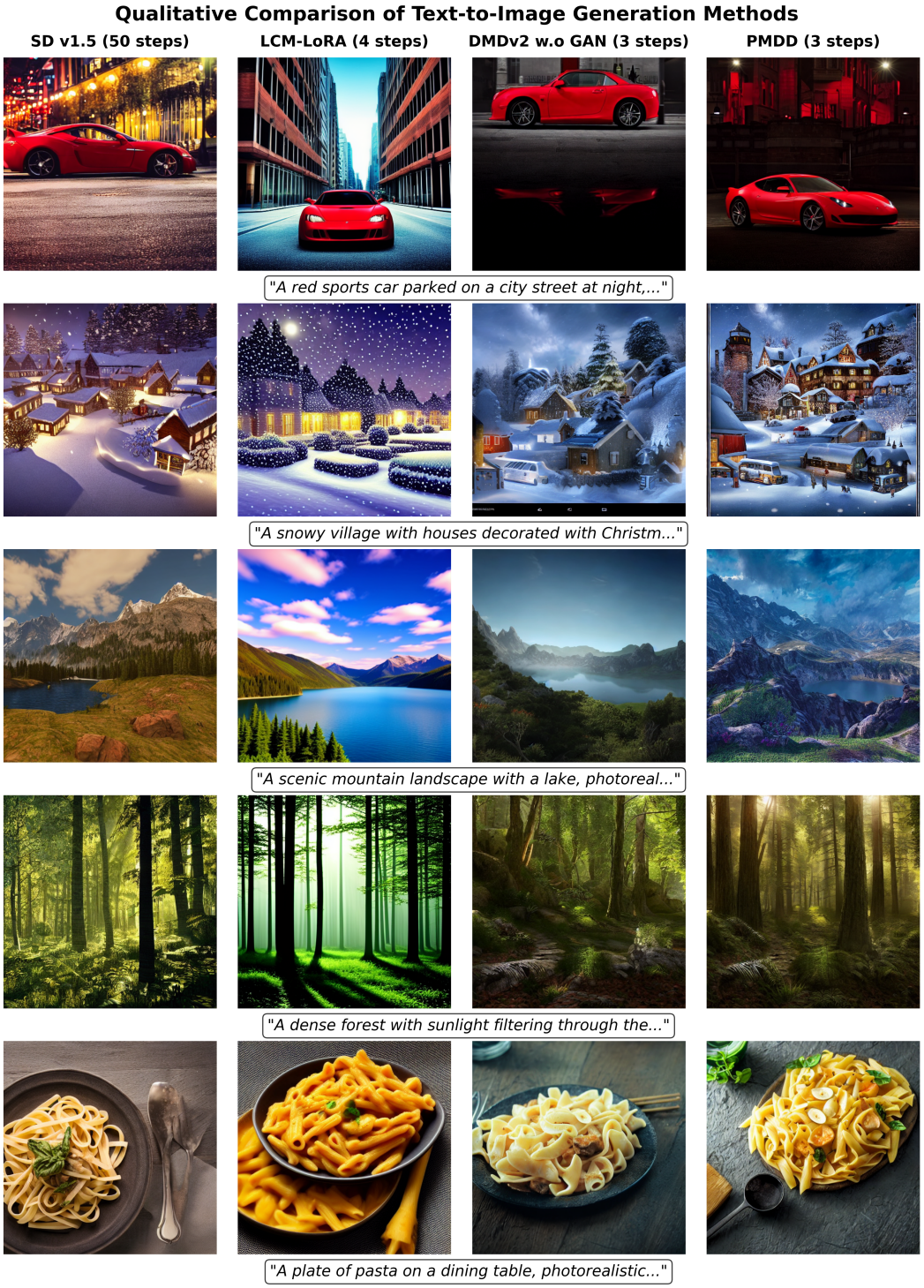

Figure 4: Comparison of text-to-image generation across Stable Diffusion v1.5 (50 steps) and other multistep diffusion distillation methods such as LCM-LoRA, PMDD, and DMD v2. Our model (final column) attains superior quality compared to other methods, with comparable or faster inference speed.

	Ext. Sampling ×		Ext. Sampling ✓	
Unforget Weight	FID (2 steps)	FID (1 step)	FID (2 steps)	FID (1 step)
$\lambda = 0.3$	8.63	11.67	8.57	11.81
$\lambda = 1.0$	8.79	11.42	8.66	11.52
$\lambda = 3.0$	8.89	10.98	8.61	10.50
$\lambda = 10.0$	9.44	10.07	9.00	9.99

Table 6: Ablation of unforget weight and external sampling of x_{t_i} on 2-step inference for COCO 2014 trained in 16K iterations.

Solvers (SEEDS)(Gonzalez et al., 2024) employs an exponential time-differencing approach separating linear terms for analytical evaluation, while SA-Solver (Xue et al., 2024) applies Adams-Bashforth integrator which controls noise injection via hyper-parameter τ. In general, diffusion samplers utilizing enhanced SDE solvers tend to be slower than those based on high-order ODE solvers (Lu et al., 2022a;b; Zheng et al., 2023b), reasoned by ODE's deterministic nature simplifying the denoising process. High-order ODE solvers typically exploit the special structures of the diffusion generation process. (Liu et al., 2022) designs the VP ODE semi-linear structure, while (Zhang & Chen, 2022; Lu et al., 2022a) further expand this concept and utilize an exponential integrator method to simplify the process. Notably, UniPC (Zhao et al., 2024), which integrates a corrector into DPM-Solver++ Lu et al. (2022b), unifies various existing methods under a predictor-corrector framework.

An alternative approach focuses on aligning the distributions of the student and teacher across different time steps. SwiftBrush (Nguyen & Tran, 2024) adapts 3D distribution matching techniques from Score Distillation Sampling (Poole et al., 2023) and Variational Score Distillation (Wang et al., 2024) to 2D image synthesis by replacing the 3D NeRF rendering component with a 2D text-to-image generator. Yin et al. (2024b) further leverages this framework by incorporating an extra regression loss for better generation capabilities. Zhou et al. (2024) generalizes this idea by replacing the reverse KL-Divergence used in original work with Fisher Divergence, featuring DMD as its special case and achieving a more general framework for student-teacher distribution alignment. A concurrent work - Zhou et al. (2025) leverages this framework to extend to multistep data-free sampling by jointly training N steps simultaneously with a single adapted network $x_\varphi(x_t, t)$ to approximate $g_\phi(x_{t_i}, t_i)$ where $x_t = a_t g_\phi(x_{t_i}, t_i) + \sigma_t \epsilon$ for all t_i.

Through extensive experiments against DMDv2 (Yin et al., 2024a) and (Zhou et al., 2025), we find that relying on a single adapted network is insufficient. In contrast, our method introduces a progressive training mechanism, employing a separate adapted network $x_\varphi(x_t, t)$ for each t_i. This strategy along with the unforget loss $\mathcal{L}_{\text{SiD}}(\phi)$ achieves superior performance compared to both Yin et al. (2024a) and Zhou et al. (2025).

6 Conclusion

In conclusion, our progressive multi-step diffusion distillation framework effectively overcomes the limitations of prior one-step and distributional-matching approaches, achieving high-fidelity generation with significantly reduced computational cost. By introducing data-free intermediate sampling and an unforget loss, our method preserves generation quality across iterations and enables efficient few-step sampling. Experimental results demonstrate that PMDD consistently outperforms existing distillation methods and even teacher models in some cases, setting a new state-of-the-art in multi-step data-free diffusion distillation while requiring far fewer resources.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=NsMLjcFaO8O.

Quan Dao, Hao Phung, Trung Tuan Dao, Dimitris N. Metaxas, and Anh Tran. Self-corrected flow distillation for consistent one-step and few-step text-to-image generation. *CoRR*, abs/2412.16906, 2024. doi: 10.48550/ARXIV.2412.16906. URL https://doi.org/10.48550/arXiv.2412.16906.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Martin Gonzalez, Nelson Fernandez Pinto, Thuy Tran, Hatem Hajri, Nader Masmoudi, et al. Seeds: Exponential sde solvers for fast high-quality sampling from diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML Workshop*, 2023.

Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/35b5c175e139bff5f22a5361270fce87-Abstract-Conference.html.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ymjI8feDTD.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, MIT, NYU, 2009. CIFAR10 and CIFAR100 were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.

Yujia Li, Kevin Swersky, and Richard S. Zemel. Generative moment matching networks. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1718–1727. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/li15.html.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=CNicRIVIPA.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *CoRR*, abs/2311.05556, 2023b. doi: 10.48550/ARXIV.2311.05556. URL https://doi.org/10.48550/arXiv.2311.05556.

Weijian Luo, Colin Zhang, Debing Zhang, and Zhengyang Geng. Diff-instruct*: Towards human-preferred one-step text-to-image generative models. *CoRR*, abs/2410.20898, 2024. doi: 10.48550/ARXIV.2410.20898. URL https://doi.org/10.48550/arXiv.2410.20898.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, pp. 14297–14306, 2023.

Thuan Hoang Nguyen and Anh Tran. 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*, pp. 7807–7816, 2024.

Viet Nguyen, Anh Nguyen, Trung Tuan Dao, Khoi Nguyen, Cuong Pham, Toan Tran, and Anh Tran. SNOOPI: supercharged one-step diffusion distillation with proper guidance. *CoRR*, abs/2412.02687, 2024. doi: 10.48550/ARXIV.2412.02687. URL https://doi.org/10.48550/arXiv.2412.02687.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. *CoRR*, abs/2406.04103, 2024. doi: 10.48550/ARXIV.2406.04103. URL https://doi.org/10.48550/arXiv.2406.04103.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 574–584. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/song20a.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Sirui Xie, Zhisheng Xiao, Diederik P Kingma, Tingbo Hou, Ying Nian Wu, Kevin Patrick Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *arXiv preprint arXiv:2405.16852*, 2024.

Chen Xu, Tianhui Song, Weixin Feng, Xubin Li, Tiezheng Ge, Bo Zheng, and Limin Wang. Accelerating image generation with sub-path linear approximation model. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LIII*, volume 15111 of *Lecture Notes in Computer Science*, pp. 323–339. Springer, 2024. doi: 10.1007/978-3-031-73668-1_19. URL https://doi.org/10.1007/978-3-031-73668-1_19.

Shuchen Xue, Mingyang Yi, Weijian Luo, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8f9d1362a386e80a723e98451a8c7564-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42390–42402. PMLR, 2023a. URL https://proceedings.mlr.press/v202/zheng23d.html.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36: 55502–55542, 2023b.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

Mingyuan Zhou, Yi Gu, and Zhendong Wang. Few-step diffusion via score identity distillation. *CoRR*, abs/2505.12674, 2025. doi: 10.48550/ARXIV.2505.12674. URL https://doi.org/10.48550/arXiv.2505.12674.