Inference-Time Diffusion Model Distillation

Geon Yeong Park[1] Sang Wan Lee[2,3*] Jong Chul Ye[1,3*]
[1]Bio and Brain Engineering, [2]Brain and Cognitive Sciences, [3]Kim Jaechul Graduate School of AI
Korea Advanced Institute of Science and Technology (KAIST)
{pky3436, sangwan, jong.ye}@kaist.ac.kr

Figure 1. Representative results comparing baselines (LCM-LoRA [29], SDXL-Lightning [21], DMD2 [47]) and the proposed framework. Results of each baseline are generated with 4 step sampling. We improve the visual fidelity and textual alignment of student models by conducting a *inference-time* diffusion distillation with the guidance of teacher model, e.g. SDXL [34], in early sampling stages (first 1 step).

Abstract

Diffusion distillation models effectively accelerate reverse sampling by compressing the process into fewer steps. However, these models still exhibit a performance gap compared to their pre-trained diffusion model counterparts, exacerbated by distribution shifts and accumulated errors during multi-step sampling. To address this, we introduce Distillation++, a novel inference-time distillation framework that reduces this gap by incorporating teacher-guided refinement during sampling. Inspired by recent advances in conditional sampling, our approach recasts student model sampling as a proximal optimization problem with a score distillation sampling loss (SDS). To this end, we integrate distillation optimization during reverse sampling, which can be viewed as teacher guidance that drives student sampling trajectory towards the clean manifold using pre-trained diffusion models. Thus, Distillation++ improves the denoising process in real-time without additional source data or fine-tuning. Distillation++ demonstrates substantial improvements over state-of-the-art distillation baselines, particularly in early sampling stages, positioning itself as a robust guided sampling process crafted for diffusion distillation models.

1. Introduction

Diffusion models have significantly advanced image generation by producing high-quality samples through an iterative refinement process that gradually denoises an initial noise vector. This refinement can be viewed as solving the reverse generative Stochastic Differential Equation (SDE) or Ordinary Differential Equations (ODE), a counterpart to a prescribed forward SDE/ODE.

Despite achieving unprecedented realism and diversity, diffusion models face a critical challenge: slow sampling speed. The progressive denoising process is computationally expensive because solving the reverse SDE/ODE requires fine discretization of time steps to minimize discretization errors considering the curvature of the diffusion sampling trajectory [15]. This leads to an increased number of function evaluations (NFE), where typical diffusion sampling necessitates tens to hundreds of NFE, limiting its use in user-interactive creative tools.

To address these limitations, various works have proposed to accelerate diffusion sampling. One promising avenue is distillation models, which distill the pre-trained diffusion models (teacher model) by directly estimating the integral along the Probability Flow ODE (PF-ODE) trajectory [41]. This effectively amortizes the computational cost of sampling into the training phase. Recent advances in distillation methods have led to the emergence of a one-step image generator; however, few-step distillation models (student models) are also often preferred in terms of image quality.

Despite progress, distillation models still face challenges, particularly in bridging the performance gap between the teacher model and its distilled student counterpart. The primary issues include potential suffer from accumulated errors in multi-step sampling or iterative training. For example, [16] identifies potential issues of multi-step sampling with models estimating the zero-time endpoint of PF-ODE. As an empirical demonstration, they show that the generation quality of consistency models does not improve as NFE increases. Similarly, [9] shows that consistency errors can accumulate across time intervals, leading to unstable training. [8, 47] warn the general training/inference mismatch observed in the multi-step sampling of student models. While prior works aim to mitigate this gap by introducing real training datasets [38, 47], it may face a potential distribution shift between datasets of teacher and student, leading to suboptimal performance on out-of-distribution (OOD) prompts.

In this work, we aim to overcome this fundamental gap between teacher and student by proposing a novel *inference-time* distillation framework called *Distillation++*. Distillation++ is a novel symbiotic distillation framework that distills the teacher model to a student model throughout the *sampling* process, in contrast to prior works which distill only during *training* process. In particular, inspired by the recent advances in text-conditional sampling [4, 5, 17], we first recast the diffusion sampling of student models to the proximal optimization problem and regularize its sampling path with a score distillation sampling loss (SDS, [36]) which opens an important opportunity for external guidance. Based on this insight, we develop a teacher-guided sampling process that inherently minimizes the SDS loss by leveraging the pre-trained diffusion model as a teacher evaluating student's denoised estimates during sampling. Specifically, intermediate estimates of student models are refined towards the clean manifold by minimizing the SDS loss, computed using the teacher model.

This inference-time distillation fosters a life-long partnership between teacher and student, allowing continuous teacher guidance beyond the training phase. Our empirical results demonstrate that this data-free distillation approach significantly improves student model performance, particularly in the early stages of sampling with minimal computational costs. We believe this approach introduces a new opportunity for inference-time distillation, a concept that has not been previously explored.

Our contributions can be summarized as follows:

- We introduce Distillation++, a novel inference-time distillation framework where teacher guides sampling process so that it closes the gap between student and teacher during sampling with affordable computational costs.
- The proposed framework is generally compatible with various student models, ranging from ones directly predicting the PF-ODE endpoint [28, 43, 48] to the progressive distillation branches [21, 37]. We also demonstrate its general applicability with various solvers, including Euler and DPM++ 2S Ancestral [26].
- To the best of our knowledge, Distillation++ is a first tuning-free and data-free inference-time distillation framework that serves as a viable post-training option for improving distillation model sampling.

2. Background

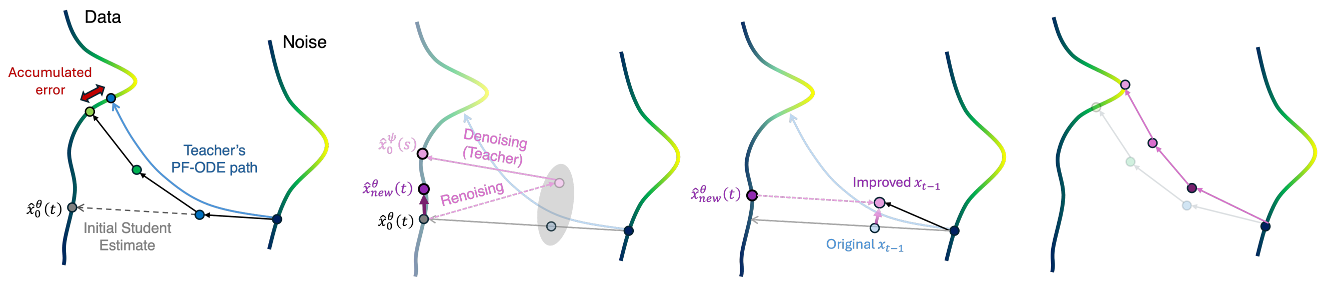
Diffusion models. Diffusion models aim to generate samples by learning the reversal of a prescribed diffusion forward process. In discrete setting with a total of N noise scales, define the fixed forward diffusion kernel as follows:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\beta_t\boldsymbol{x}_{t-1}, (1-\beta_t)I), \tag{1}$$

$$p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}}\boldsymbol{x}_0, (1-\bar{\alpha})I), \tag{2}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^d \sim p_0(\boldsymbol{x})$ is given as a clean sample, β_t denotes a noise schedule discretized from $\beta(t) : \mathbb{R} \to \mathbb{R} > 0$, $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \Pi_{i=1}^t \alpha_i$. Then as $N \to \infty$, the underlying forward noising process can be expressed as the forward Itô SDE [42] given $\boldsymbol{x}(t) \in \mathbb{R}^d$:

$$d\boldsymbol{x} = -\frac{\beta(t)}{2}\boldsymbol{x}dt + \sqrt{\beta(t)}d\boldsymbol{w}, \tag{3}$$



(a) Denoised estimates (b) Inference-time Distillation (c) Renoising (d) Improved trajectory

Figure 2. **Overview.** (**a**) Diffusion models (in blue) sample by solving the PF-ODE, requiring a computationally expensive integral from time T to 0. Student models (in black) accelerate sampling by approximating this integral, but their (initial) estimates are often suboptimal. (**b**) To bridge this gap post-training, we propose an *inference-time* distillation. Specifically, we refine the student models' initial estimates by refining them towards teacher estimates, obtained by consecutive renoising and denoising, as in (9). (**c**) This process steers the sampling trajectory closer to the teacher model's distribution, thereby (**d**) improving the sampling path.

where $\boldsymbol{w}$ is the d-dimensional standard Wiener process. Then the counterpart reverse SDE is defined as follows [1]:

$$d\boldsymbol{x} = \Big[-\frac{\beta(t)}{2}\boldsymbol{x} - \beta(t)\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\Big]dt + \sqrt{\beta(t)}d\bar{\boldsymbol{w}},$$

where $\bar{\boldsymbol{w}}$ is the d-dimensional standard backward Wiener process. The deterministic counterpart PF-ODE [42] can be similarly defined. Then the goal of diffusion model training is to approximate a score function $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ by denoising score matching (DSM):

$$\min_\theta \mathbb{E}_{\boldsymbol{x}(t),\boldsymbol{x}_0,t}\big[\left\|\boldsymbol{s}_\theta(\boldsymbol{x}(t),t) - \nabla_{\boldsymbol{x}(t)} \log p_t(\boldsymbol{x}(t)|\boldsymbol{x}_0)\right\|\big],$$

where $\boldsymbol{x}_0 \sim p_0(\boldsymbol{x})$ denotes a clean sample. It can be shown that this score matching is equivalent to the epsilon matching with different parameterization of residual denoiser $\boldsymbol{\epsilon}_\theta$:

$$\min_\theta \mathbb{E}_{\boldsymbol{x}(t),\boldsymbol{x}_0,\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\big[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}(t),t) - \boldsymbol{\epsilon}\|\big], \tag{4}$$

and $\boldsymbol{s}_{\theta^*}(\boldsymbol{x}(t),t) \simeq -\frac{\boldsymbol{x}(t)-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1-\bar{\alpha}} = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\theta^*}(\boldsymbol{x}(t),t)$.

Based on this, generative sampling can be performed by solving the PF-ODE, equivalent to computing the integral from time T to 0, approximated by various off-the-shelf ODE solvers. For instance, a single iteration of DDIM sampling [40] reads:

$$\hat{\boldsymbol{x}}_0^\theta(t) = (\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t))/\sqrt{\bar{\alpha}_t} \tag{5}$$

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_0^\theta(t) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)), \tag{6}$$

where $\boldsymbol{x}_t \in \mathcal{M}_t$ with noisy manifold $\mathcal{M}_t$, and $\hat{\boldsymbol{x}}_0(t) = \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ is the denoised estimate, which can be equivalently derived using Tweedie's formula [6].

For a text-conditional sampling with text embedding $\boldsymbol{c}$, the classifier-free guidance (CFG, [12]) is widely leveraged:

$$\hat{\boldsymbol{\epsilon}}_\theta^w(\boldsymbol{x}_t,t,\boldsymbol{c}) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\varnothing) + w(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{c}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\varnothing)),$$

where $\boldsymbol{c} = \varnothing$ refers to the null text embedding. For notational simplicity, we will interchangably use $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\varnothing), \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{c}), \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)$ and similarly $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{c}), \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{c})$ unless ambiguity arises.

Diffusion acceleration. While diffusion models generate high-quality samples with a relatively stable training procedure of (4), the numerical integration of the PF-ODE is computationally expensive. Two primary strategies have been developed to accelerate this process. The first strategy corresponds to the fast diffusion samplers which improves time-discretized numerical integration methods [14, 15, 25, 40]. Despite its interesting progress, further reducing NFE can significantly degrade performance in practice.

Alternatively, diffusion distillation has emerged as a promising approach to accelerate the sampling by indirectly estimating the entire integral of PF-ODE. For instance, [27] trains the denoising student that predicts the endpoint of the PF-ODE given an initial noise vector. Similarly, [48] trains a one-step generator by matching the pre-compute diffusion outputs with using the distribution matching training objective. [2, 10, 43] further learn to map a different point on the ODE trajectory to its endpoint or to the boundaries of sub-intervals. In a related yet distinct approach, progressive distillation methods have been developed considering the expensive computational cost of simulating the full denoising trajectory for each loss function evaluation. These methods often iteratively train a series of student models, each halving the number of sampling steps required by the previous model.

3. Main Contribution: Distillation++

While both improved sampling and distillation methods have made significant progress in addressing the speed-quality trade-off, there has been limited advances in integrating these methods through the design of specialized solvers for improving multi-step distillation models. This can be attributed to two primary reasons. First, off-the-shelf ODE solvers typically require the estimation of the tangent gradient direction along the solution trajectory. In contrast, many distillation models [28, 43, 47, 48] directly predict the endpoint of the trajectory on the side of data distribution at any given time point, avoiding the need for direct trajectory estimation. Consequently, the multi-step sampling procedures of these student models are reduced to simple iterative processes involving random noise injection and subsequent denoising steps [43]. Furthermore, distillation models often sample only 2-8 steps, which may restrict the design space for solvers, including higher-order ones.

These constraints limit post-training options for improving distillation model sampling, despite the performance gap between multi-step student and teacher models. To address this, we propose *Distillation++*, a novel inference-time distillation framework. Specifically, Distillation++ leverages large-scale pre-trained diffusion models as a teacher signal during the early-stage sampling process (e.g. first 1-2 steps), which substantially improves the overall sampling trajectory as shown in Fig. 2.

3.1. Derivations of Distillation++

For a better multi-step sampling of student models, we derive a novel data-free distillation process by integrating the teacher guidance as an optimization problem within the reverse sampling process. Let $\theta \in \mathbb{R}^D$ and $\psi \in \mathbb{R}^D$ parameterize the residual denoiser of student and teacher diffusion models, respectively. Then, our objective is to define a guidance loss function ℓ_{distill} that, when minimized under the symbiotic guidance of the teacher model, progressively aligns the *student's* intermediate estimates $\hat{\boldsymbol{x}}_0^\theta(t)$ with the teacher model's distribution. We present such guidance loss function ℓ_{distill} as a score distillation sampling loss (SDS) with respect to the *teacher model* (ψ) as follow:

$$\begin{aligned}\ell_{\text{distill}}(\boldsymbol{x};\psi,s) &= \left\| \boldsymbol{\epsilon}_\psi(\underbrace{\sqrt{\bar{\alpha}_s}\boldsymbol{x} + \sqrt{1-\bar{\alpha}_s}\boldsymbol{\epsilon}}_{\boldsymbol{x}_s}, \varnothing) - \boldsymbol{\epsilon} \right\|_2^2 \\ &= \left\| \frac{\boldsymbol{x}_s - \sqrt{\bar{\alpha}_s}\hat{\boldsymbol{x}}_0^\psi(s)}{\sqrt{1-\bar{\alpha}_s}} - \frac{\boldsymbol{x}_s - \sqrt{\bar{\alpha}_s}\boldsymbol{x}}{\sqrt{1-\bar{\alpha}_s}} \right\|_2^2 \\ &= \frac{\bar{\alpha}_s}{1-\bar{\alpha}_s} \left\| \boldsymbol{x} - \hat{\boldsymbol{x}}_0^\psi(s) \right\|_2^2, \end{aligned} \tag{7}$$

where $\boldsymbol{x}_s = \sqrt{\bar{\alpha}_s}\boldsymbol{x} + \sqrt{1-\bar{\alpha}_s}\boldsymbol{\epsilon}$ with a perturbation time $s > 0$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, I), \boldsymbol{x} \in \mathcal{M}$ with a clean data manifold $\mathcal{M}$, and $\hat{\boldsymbol{x}}_0^\psi(s)$ follows the Tweedie's formula in (5) which is denoised by the teacher model ψ. For simplicity, we consider the text-unconditional version with the null-text embedding $\varnothing \in \mathbb{R}^{d'}$ in this subsection. This loss represents an ideal condition that high-quality denoised estimates should satisfy: the ideal student samples should be well reconstructed from random perturbations followed by denoising using large-scale pre-trained teacher diffusion models. Variants of the SDS framework [30, 32, 48] have thus frequently been employed as key components in recent diffusion distillation training procedure.

Then, we can now integrate the optimization step of ℓ_{distill} in terms of denoised student estimates $\hat{\boldsymbol{x}}_0^\theta(t)$, resulting DDIM sampling process (5). A potential concern includes the feasibility of gradient descent due to the intractable Jacobian computation: $\frac{d\ell}{d\boldsymbol{x}} = \frac{d\ell}{d\boldsymbol{\epsilon}_\psi}\frac{d\boldsymbol{\epsilon}_\psi}{d\boldsymbol{x}}$. To circumvent this, by following the prior work on Decomposed Diffusion Sampling (DDS) [4], which bypasses direct computation of the score Jacobian, we have

$$\begin{aligned}\boldsymbol{x}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\Big(\hat{\boldsymbol{x}}_0^\theta(t) - \gamma_t \nabla_{\hat{\boldsymbol{x}}_0^\theta(t)} \ell_{\text{distill}}\big(\hat{\boldsymbol{x}}_0^\theta(t);\psi,s\big)\Big) \\ &\quad + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t),\end{aligned} \tag{8}$$

where $\gamma_t > 0$ refers to the step size. [3, 5] supports that this update allows precise transition to the subsequent noisy manifold $\mathcal{M}_{t-1}$ under some manifold assumption. This gives us a simple single DDIM sampling iterate as follows:

$$\begin{aligned}\hat{\boldsymbol{x}}_{\text{new}}^\theta(t) &= \hat{\boldsymbol{x}}_0^\theta(t) - \lambda\big(\hat{\boldsymbol{x}}_0^\theta(t) - \hat{\boldsymbol{x}}_0^\psi(s)\big) \\ &= (1-\lambda)\hat{\boldsymbol{x}}_0^\theta(t) + \lambda\hat{\boldsymbol{x}}_0^\psi(s), \\ \boldsymbol{x}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_{\text{new}}^\theta(t) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t),\end{aligned} \tag{9}$$

where $\lambda = \frac{2\gamma_t\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}$. Note that the updated estimate $\hat{\boldsymbol{x}}_{\text{new}}^\theta(t)$ can be obtained by the interpolation between initial student estimate $\hat{\boldsymbol{x}}_0^\theta(t)$ and the revised teacher estimate $\hat{\boldsymbol{x}}_0^\psi(s)$. Thus, by taking the gradient step of guidance loss, we can update $\hat{\boldsymbol{x}}_0^\theta(t)$ towards the clean manifold well-aligned with the teacher model distribution.

Renoising strategy. We empirically observed that the time step schedule of s plays a crucial role in performance, consistent with recent score distillation frameworks. Considering that the teacher model is well-trained on fine level of timesteps, it is compatible with a broad range of renoising timestep s. That said, as our approach recasts distillation more as a guided sampling rather than a mere training process, we adopt a decreasing time step schedule for s as $s = t-1$ following the sampling process (Fig. 2b), reminiscent of schedules used in [17, 45, 50]. This is in contrast to conventional random timestep s scheduling [36]. Intuitively, as the student model often learns to leap towards the end-point of each sub-interval (Fig. 2a), refining this large-step update direction at the terminal of each sub-interval with teacher model may better guide the sampling trajectory (Fig. 2c,d). Empirical evidences and more discussions are provided in Table 4.

Teacher guidance. For a intuitive understanding of Distillation++, we provide reparameterization of (9), by intentionally assuming $\bar{\alpha}_t \approx \bar{\alpha}_s$. This leads to the formulation of *teacher guidance*, drawing parallels to the CFG denoising mechanism:

$$\hat{\boldsymbol{x}}^{\theta}_{\text{new}}(t) = \frac{\tilde{\boldsymbol{x}}_t - \sqrt{1-\bar{\alpha}_t}\big(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t) + \lambda\big(\boldsymbol{\epsilon}_\psi(\boldsymbol{x}_s,s) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)\big)}{\sqrt{\bar{\alpha}_t}}$$
$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}^{\theta}_{\text{new}}(t) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t), \tag{10}$$

where $\lambda > 0$ serves as a guidance scale, and $\tilde{\boldsymbol{x}}_t = (1-\lambda)\boldsymbol{x}_t + \lambda\boldsymbol{x}_s$. We conduct experiments using (9) in practice as (10) serves primarily as an approximation for intuition; however, (9) and (10) reproduce empirically similar generations with negligible differences, particularly with sampling steps ≥ 4 (empirical comparisons in Fig. 7). This reframework suggests that the inference-time distillation serves as a directional guidance $\lambda(\boldsymbol{\epsilon}_\psi(\boldsymbol{x}_s,s) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t))$, which steers the sampling trajectory closer to the teacher model.

Remarks. We find that applying inference-time distillation at only a small number of initial sampling steps (e.g., 1-2 steps) already enhances the overall sampling procedure. This is because each student sampling step compresses hundreds of steps of the teacher model, maximizing the guidance effects with minimal extra computations. Moreover, since students typically reuses the teacher's VAE, it further minimizes additional memory overhead. Attributed to these considerations, the peak memory of our framework remains $\leq$ 24GB on a standard GPU as in Table 3.

Importantly, Distillation++ integrates seamlessly with various distillation models, especially those directly predicting the PF-ODE endpoint, such as consistency models [43] , which are not amenable to ODE solvers like DDIM in Sec. 3.1. While Sec. 3.1 simplifies by focusing on a specific solver, Distillation++ is generalizable, even with simple iterative processes involving denoising followed by stochastic renoising steps, as in multi-step consistency models.

3.2. Text-conditional sampling and other solvers

Section 3.1 focused on the text-unconditional setting for simplicity. Here we incorporate the teacher guidance with widespread classifier-free guidance (CFG, [12]) for text-conditional sampling. Let $\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(t)$ denote the text-conditioned clean estimate defined as:

$$\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(t) = \frac{\tilde{\boldsymbol{x}}_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}^{w}_{\theta}(\boldsymbol{x}_t,\boldsymbol{c})}{\sqrt{\bar{\alpha}_t}}, \tag{11}$$

where we will interchangeably use $\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(\boldsymbol{x}_t)$ and similarly define $\hat{\boldsymbol{x}}^{\psi}_{\boldsymbol{c}}(t)$. Then, based on (9), the resulting text-conditional version of $\hat{\boldsymbol{x}}^{\theta}_{\text{new}}(t)$ in (9) is derived as follows: $\hat{\boldsymbol{x}}^{\theta}_{\text{new},\boldsymbol{c}}(t) = (1-\lambda)\hat{\boldsymbol{x}}^{\theta}_{\boldsymbol{c}}(t) + \lambda\hat{\boldsymbol{x}}^{\psi}_{\boldsymbol{c}}(s)$. Here, $\hat{\boldsymbol{x}}^{\theta}_{\text{new},\boldsymbol{c}}(t)$ can be similarly derived in the form of teacher guidance as in (10) with the same assumption on $\bar{\alpha}_t$:

$$\frac{\tilde{\boldsymbol{x}}_t - \sqrt{1-\bar{\alpha}_t}\big(\hat{\epsilon}^{w}_{\theta}(\boldsymbol{x}_t,\boldsymbol{c}) + \lambda\big(\hat{\epsilon}^{w}_{\psi}(\boldsymbol{x}_s,\boldsymbol{c}) - \hat{\epsilon}^{w}_{\theta}(\boldsymbol{x}_t,\boldsymbol{c})\big)}{\sqrt{\bar{\alpha}_t}}.$$

The overall pipeline is outlined in algorithm 1.

While Section 3.1 is delineated with DDIM or equivalently Euler solvers, the underlying principle can be shared with other conventional solvers, e.g. Karras [15], PNDM [22], etc. Please refer to the appendix for derivations.

4. Experimental results

4.1. Implementation details

In our experiments, we demonstrate the impacts of distillation++ using SDXL backbone [34] and its open-sourced weights. All experiments are conducted using a single NVIDIA GeForce RTX 4090. All quantitative results are obtained with one-step teacher-guided distillation at the initial sampling step of the student models (i.e., $t = T$), minimizing additional computational costs, though more frequent guidance could further enhance quality. Detailed implementation settings are provided in the appendix and the following.

Baselines. Distillation++ can be applied to various state-of-the-art T2I distillation baselines which are widely used with open-sourced weights. LCM [28] is a consistency model that operates in the image latent space of a pretrained autoencoder, solving an augmented PF-ODE in this space. LCM-LoRA [29] extends LCM by distilling pre-trained latent diffusion models into LoRA parameters, significantly reducing memory requirements while maintaining high generation quality. DMD2 [47] improves distribution-matching distillation (DMD, [48]) by enabling multi-step sampling and eliminating the needs for expensive regression loss. SDXL-Turbo [38] integrates both adversarial loss and score-distillation sampling loss. Building on similar adversarial training regime, SDXL-Lightning [21] inherits progressive distillation [37], relaxing mode coverage and also includes a LoRA variant (SDXL-Lightning LoRA).

Setup. For visual fidelity, we evaluate Fréchet Inception Distance (FID, [11]) of images generated from the random 10K prompts of MS-COCO validation set. For text-alignment and user-preference, we additionally measure ImageReward [46] and PickScore [18] which are more reliable compared to the existing text-image scoring metrics, such as CLIP similarity. For Table 1, we use 4 step Euler sampling for SDXL-Lightning and its LoRA variant, 4 step iterative random sampling for DMD2, LCM, and LCM-LoRA (incompatible with conventional solvers), and 6 step DPM++ 2S Ancestral sampling [26] for SDXL-Turbo, utilizing DreamShaper [31], an open-source customized model from the community.

4.2. Evaluation and analysis

Quantitative Analyses. As shown in Table 1, Distillation++ consistently improves visual quality, text alignment, and user preference across various distillation model baselines and solvers, with only a single additional step of teacher (SDXL) evaluation. Most multi-step distillation models offers only a preset number of sampling steps with a fixed time-step schedule, limiting sampling flexibility. Distillation++ ad-

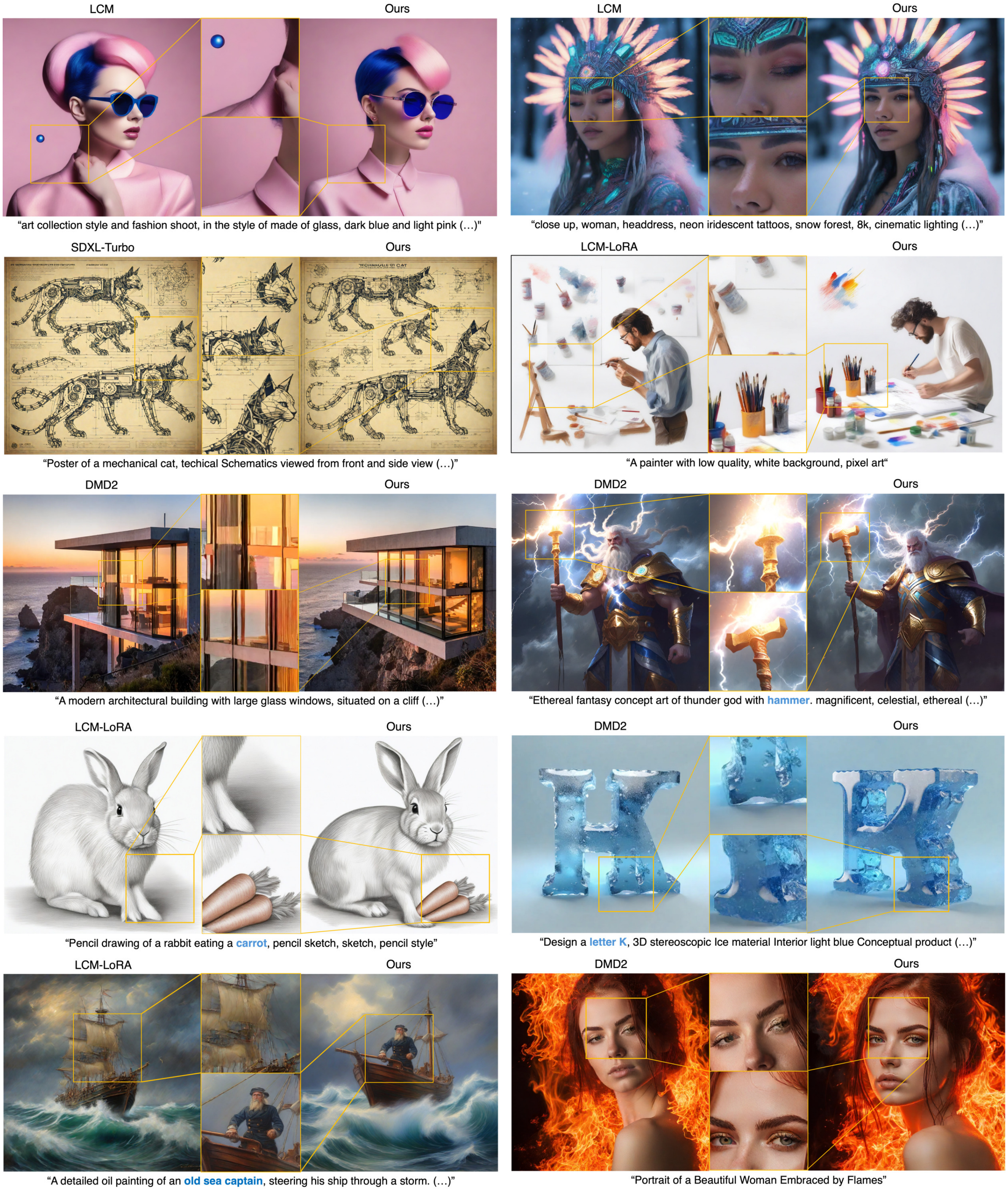


Figure 3. Qualitative comparisons against state-of-the-art distillation baselines. Baselines using 4 sampling steps: SDXL-Lightning, DMD2, SDXL-Turbo. Baselines using 8 sampling steps: LCM, LCM-LoRA. By conducting the inferece-time distillation in early sampling stages, we reduce artifacts, improve the visual fidelity and textual alignment.

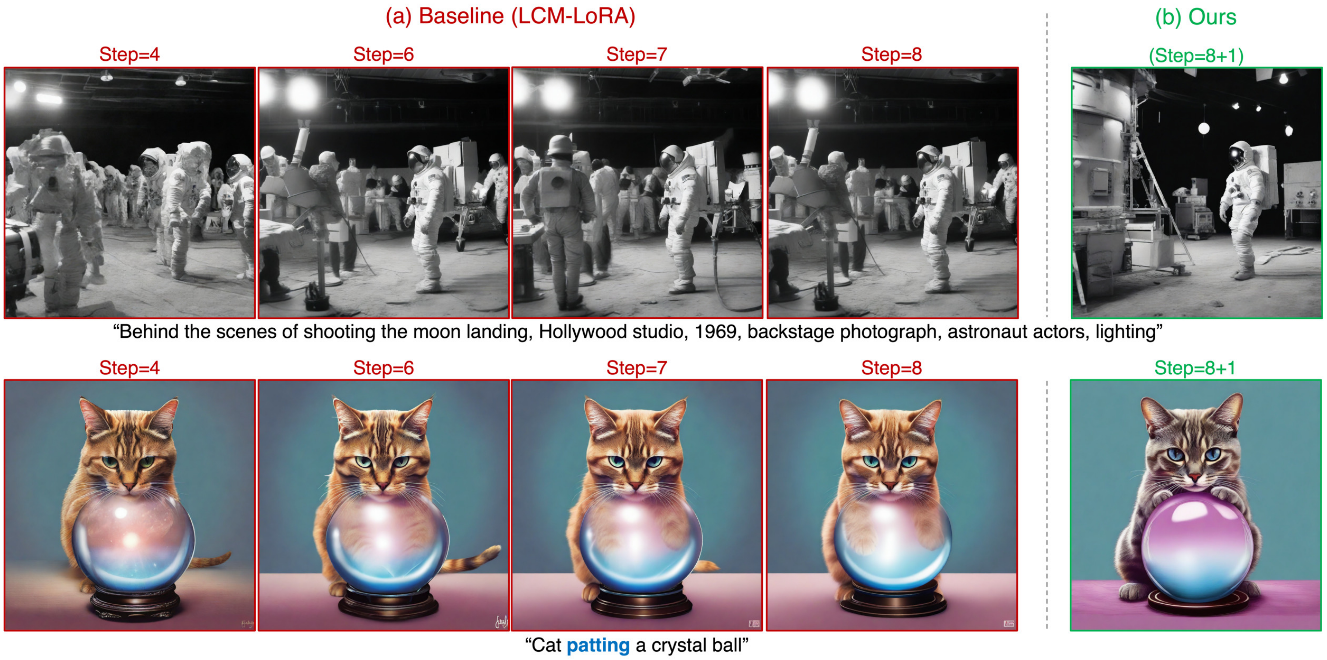


Figure 4. (**a**) Results of baseline (LCM-LoRA) with varying number of sampling steps (4, 6, 7, 8). Increasing the number of sampling steps of student models does not guarantee improvements in textual alignment or physical feasibility. (**b**) Our improved results with inference-time distillation. Teacher guidance is applied only at the first of 8 steps (total step=8+1).

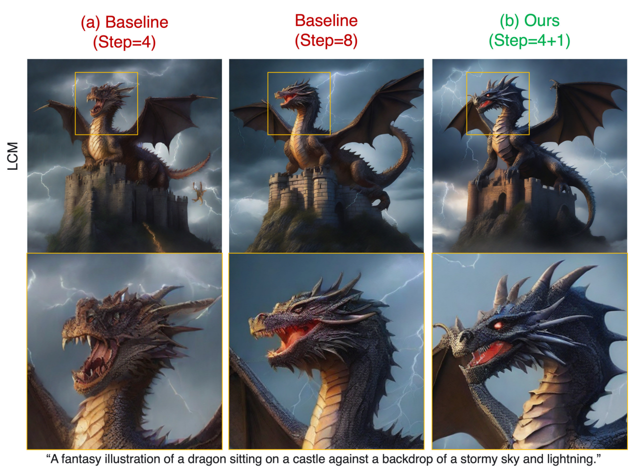


Figure 5. (**a**) Results of baseline (LCM) with 4 and 8 sampling steps. (**b**) Ours with 4 step sampling + 1 step distillation.

Table 1. Quantitative evaluation on MS-COCO 10K with 4 step baseline sampling and 1 step additional inference-time distillation. Light refers to SDXL-Lightning.

Models	FID (↓)	ImageReward (↑)	PickScore (↑)
LCM [28]	20.674	0.561	0.494
LCM++	**20.149**	**0.597**	**0.505**
LCM-LoRA [29]	20.300	0.494	0.490
LCM-LoRA++	**19.815**	**0.522**	**0.510**
Light [21]	24.506	0.787	0.496
Light++	**23.876**	**0.820**	**0.503**
Light-LoRA [21]	25.304	0.750	0.482
Light-LoRA++	**24.429**	**0.778**	**0.518**
DMD2 [47]	21.238	0.777	0.490
DMD2++	**20.937**	**0.797**	**0.510**
Turbo [38]	18.612	0.296	0.499
Turbo++	**18.481**	**0.310**	**0.501**

dresses this limitation and provide a post-training option for sampling steps, effectively reduces artifacts, improving fidelity and semantic alignment [33], as shown in Fig. 3.

Computational Costs and Performance. Distillation++ serves as an efficient sampling correction, though it introduces additional function evaluation: (1) In diffusion models, the spatial layout is largely determined in the early sampling stages, suggesting that early-stage guidance is required to rectify the physical feasibility and text adherence. (2) Increasing the sampling steps of the student may not guarantee these corrections due to the inherent gap between student and teacher models. For instance, Fig. 4 and 5 demonstrate that while additional steps may enhance fidelity, it may not be sufficient to capture the intended semantics, e.g. 'cat *patting*' or improves physical feasibility, e.g. well-structured spaceship launcher. Distillation++ mitigates this and nudges the sampling path towards a more feasible region.

Table 2 supports these observations where we have compared the wall-clock time and performance of Distillation++ with its counterpart baselines. It shows that the one-step inference-time distillation takes comparable or even shorter wall time compared to that of a single student sampling step, even though the teacher model relies on CFG. This is attributed to the parallel computing of both student and teacher models and batch-wise prediction of conditional and unconditional scores. That said, Distillation++ achieves consistent performance improvement without compromising significant computational costs as shown in Table 2.

Table 2. Comparative study on wall time and generation performance (MS-COCO 10K). 4 + 1 step refers to a 4 step student model sampling and 1 step inference-time distillation. Wall-clock time is measured per each prompt.

Metrics	LCM [28]				LCM-LoRA [29]			
	4+1 step	5 step	5+1 step	6 step	4+1 step	5 step	5+1 step	6 step
FID (↓)	**20.149**	20.732	**20.369**	21.540	**19.815**	20.579	**20.244**	21.039
ImageReward (↑)	**0.597**	0.593	**0.603**	0.585	**0.522**	0.518	**0.528**	0.519
Time (sec, ↓)	**1.987**	1.996	**2.241**	2.250	**1.828**	1.830	**2.067**	2.085

	LCM-LoRA (SD1.5)	+ Ours	SDXL-lightning	+ Ours
VRAM	3.55GB	5.30GB	10.86GB	16.10GB

Table 3. Peak GPU memory usage in sampling (SD1.5, SDXL).

Ablation study on renoising process. Table 4 highlights the importance of decreasing (reverse-diffusion) time step schedule in the renoising process (7) of Distillation++, where $s = t - \triangle t$ outperforms random timestep renoising strategy or $s = t$. This is in line with empirical findings from prior score distillation studies [17, 45, 50], which contrasts with the random timestep sampling used as standards in prior SDS works [19, 36]. We hypothesize this effect arises primarily from the progressive refinement of estimates and hierarchical minimization of KL-divergence score distillation loss, which warrants further exploration in future SDS works.

Teacher guidance. Figure 7 compares Distillation++ with interpolative denoising in (9) against teacher guided approximation in (10). This resemblance shows that the key underlying principle of Distillation++ is to modulate the denoising path under the guidance of teacher models.

Compatibility with other teacher models and multi-step distillation. Our distillation method differs fundamentally from existing self-guidance approaches by enabling collaboration between distinct diffusion models. This open new applications through diverse teacher–student pairings. e.g, using diverse fine-tuned teacher variants as in Fig 6a (Dreamshaper). Also, while we mainly verified improvements with a single-step distillation due to efficiency, additional steps may further improve results as in Fig. 6b.

5. Conclusion

This paper fosters a symbiotic collaboration between two diffusion models: fast but suboptimal student models and slower, high-quality teacher models. Distillation++ serves as a teacher-guided sampling method, minimizing SDS loss by leveraging a pre-trained model to evaluate student estimates during sampling. This approach may open avenues for exploring diffusion model ensembles or combination with flow matching models [23, 24] for synergistic sampling. Extending Distillation++ to video diffusion distillation would be a promising direction for future work, where video-domain distillation lags behind image-domain quality.

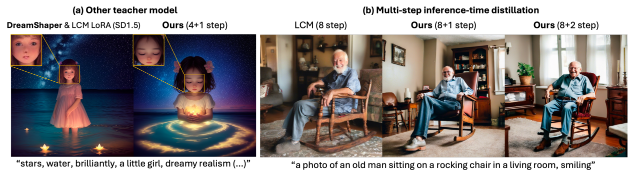


Figure 6. (**a**) Inference-time distillation with fine-tuned SD1.5 (Dreamshaper) teacher model. Student: LCM-LoRA SD1.5 (4 step). (**b**) Example of one- and two-step inference-time distillation.

Table 4. Ablation study on renoising time schedule (MS-COCO).

Models	DMD2 [47]	$s =$ random t	$s = t$	$s = t - \triangle t$
FID (↓)	21.238	21.105	21.342	**20.937**
ImageReward (↑)	0.777	0.771	0.777	**0.797**

Figure 7. Comparison on inference-time distillation using (**a**) interpolative denoising process in (9), and (**b**) teacher guidance (10).

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00233251, System3 reinforcement learning with high-level brain functions), (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)) National Research Foundation of Korea(NRF) funded by the Korean government (MSIT) (No. RS-2024-00439903, RS-2024-00341805), (No. RS-2024-00457882, AI Research Hub Project; No. RS-2025-02304967, AI Star Fellowship (KAIST)), National Research Foundation of Korea under Grant RS-2024-00336454.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 3

[2] 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. 3

[3] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. 4

[4] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. *arXiv preprint arXiv:2303.05754*, 2023. 2, 4

[5] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. 2, 4, 1

[6] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 3

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[8] Risheek Garrepalli, Shweta Mahajan, Munawar Hayat, and Fatih Porikli. Ddil: Improved diffusion distillation with imitation learning. *arXiv preprint arXiv:2410.11971*, 2024. 2

[9] Zhengyang Geng, Ashwini Pokle, William Luo, Justin Lin, and J Zico Kolter. Consistency models made easy. *arXiv preprint arXiv:2406.14548*, 2024. 2

[10] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023. 3

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

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 5

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[14] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021. 3

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 2, 3, 5, 1

[16] 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. *arXiv preprint arXiv:2310.02279*, 2023. 2

[17] Jeongsol Kim, Geon Yeong Park, and Jong Chul Ye. Dreamsampler: Unifying diffusion sampling and score distillation for image manipulation. *arXiv preprint arXiv:2403.11415*, 2024. 2, 4, 8

[18] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023. 5

[19] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 8

[20] Shanchuan Lin and Xiao Yang. Animatediff-lightning: Cross-model diffusion distillation. *arXiv preprint arXiv:2403.12706*, 2024. 2

[21] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 1, 2, 5, 7

[22] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. 5

[23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 8

[24] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022. 3

[26] 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*, 2022. 2, 5, 1

[27] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 3

[28] 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*, 2023. 2, 4, 5, 7, 8

[29] 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. *arXiv preprint arXiv:2311.05556*, 2023. 1, 5, 7, 8, 2

[30] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[31] ID: LyKon. Dreamshaper xl v2.1 turbo dpm++ sde. `https://civitai.com/models/112902/dreamshaper-xl`, 2024. Pretrained model released in CIVITAI. 5, 2

[32] 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*, pages 7807–7816, 2024. 4

[33] Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 5

[35] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[36] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2, 4, 8

[37] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 2, 5

[38] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 2, 5, 7

[39] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024. 3

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR*, 2021. 3, 1

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*, 2021. 2, 1

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 3

[43] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2, 3, 4, 5

[44] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 2

[45] Zike Wu, Pan Zhou, Xuanyu Yi, Xiaoding Yuan, and Hanwang Zhang. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9892–9902, 2024. 4, 8

[46] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[47] 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*, 2024. 1, 2, 4, 5, 7, 8

[48] 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*, pages 6613–6623, 2024. 2, 3, 4, 5

[49] Yuanhao Zhai, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, David Doermann, Junsong Yuan, and Lijuan Wang. Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. *arXiv preprint arXiv:2406.06890*, 2024. 2

[50] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *The Twelfth International Conference on Learning Representations*, 2023. 4, 8