D2PO: Optimizing Diffusion Samplers via Dynamic Preference

Jinkyu Kim*[1], Jinyoung Choi*[3], and Bohyung Han[1,2]

[1]ECE & [2]IPAI, Seoul National University, Korea
[3]AIGS, Ulsan National Institute of Science and Technology, Korea

Abstract. We propose D2PO (Dynamic Direct Preference Optimization), a principled framework for optimizing diffusion sampling policies with respect to timestep schedules and classifier-free guidance (CFG) weights. Our work is motivated by a fundamental limitation of existing student-teacher regression frameworks; low-NFE student samplers are trained to mimic high-NFE teachers, often sacrificing high-frequency texture fidelity while preserving coarse global structures, thereby misaligning the sampler with perceptual quality. D2PO addresses this challenge by reformulating sampler optimization as a preference-based alignment problem, leveraging the Direct Preference Optimization (DPO) framework. To make DPO applicable to diffusion samplers, we model the sampling policy as an energy-based model (EBM), transforming preference comparisons into tractable energy differences. We further introduce a novel energy formulation derived directly from the pretrained score network, enabling preference evaluation in perturbed spaces that jointly capture structural consistency and fine-grained details. Moreover, we introduce dynamic preferences, where the preferred samples used for alignment progressively improve as the sampling policies are learned. This self-improving mechanism replaces rigid static teacher supervision with an iterative, preference-guided refinement process, providing progressively stronger alignment signals. Extensive experiments demonstrate that D2PO aligns diffusion samplers with perceptual quality more faithfully, unlocking the full potential of high-quality teachers and consistently outperforming conventional regression-based schedulers under low-NFE constraints.

Keywords: Diffusion models · Time step optimization · Direct preference optimization

1 Introduction

Diffusion Probabilistic Models (DPMs) [16, 42, 44] have achieved unprecedented fidelity in high-resolution image synthesis, text-to-image generation [9, 36], and video generation [18,41,50,68]. However, this performance comes at a substantial

*indicates equal contribution.

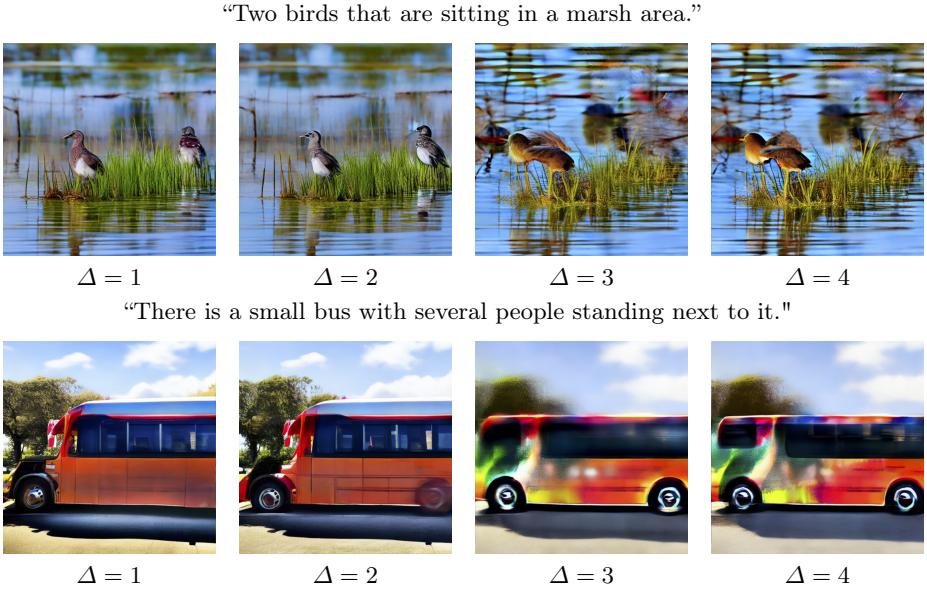

Fig. 1: Qualitative evidence of the performance bottleneck in LD3 [46].All images are generated by the same model with NFE = 4. The columns show the impact of increasing the NFE gap ($\Delta = T - S$) between the teacher (T) and the student (S). While a small gap ($\Delta = 1$) yields high-quality outputs, larger gaps (up to $\Delta = 4$) lead to severe artifacts, demonstrating LD3's inability to leverage high-fidelity teachers.

computational cost. DPMs are inherently iterative, requiring many function evaluations (NFE) during sampling, which makes high-quality generation expensive and limits practical deployment.

A broad range of approaches has been explored to mitigate this bottleneck, including accelerated numerical solvers [8,10,29,31,32,62,66], few-step knowledge distillation [22,38,39,43,59,60,67,69,70], architectural modifications [33,58], and training-time improvements [20,23,47,55]. More recently, directly optimizing the sampling policy parameters—such as timestep schedules [12, 25, 37, 46, 52, 56], classifier-free guidance weights [13], and high-order solver coefficients [12, 49, 63]—has emerged as a critical direction for acceleration.

In optimizing these sampler parameters, prior works typically adopt either a distribution or an instance perspective. Specifically, one line of work optimizes distribution-level metrics, such as FID [25] or KID [52] over large sample sets, but such population-level objectives yield weak, high-variance gradients for the low-dimensional sampler parameters. Alternatively, instance-wise distillation methods [12,46] regress a low-NFE student sampler onto the trajectories or outputs of a fixed high-NFE teacher via ℓ_2 or LPIPS [65] losses. Although effective when the student-teacher gap is moderate, this regression paradigm exhibits a structural limitation when aggressive acceleration is desired.

When the NFE gap between the student and teacher becomes substantial—a common scenario when striving for maximum acceleration—this regression-based

objective forces the student sampler to approximate a high-fidelity trajectory within its restricted capacity, compromising generation quality. This structural constraint often suppresses high-frequency textures and fine-grained details, preventing the student from fully benefiting from stronger teachers. We empirically validate this phenomenon in Fig. 1; as the teacher NFE increases while the student NFE remains fixed, the perceptual quality degrades, even for a state-of-the-art scheduler (LD3 [46]). This degradation directly reflects the structural rigidity of fixed-teacher regression, where the student is forced to prioritize low-level alignment over perceptual quality, failing to discover more effective sampling paths.

To address this limitation, we reinterpret sampler optimization as a preference-based alignment problem rather than a regression-to-teacher task. We introduce D2PO (Dynamic Direct Preference Optimization), a framework inspired by DPO but adapted to diffusion sampling policies. Applying DPO to diffusion samplers is non-trivial because the marginalized log-probability is intractable. To obtain a tractable surrogate, we model the policy-induced distribution as an Energy-Based Model (EBM). We define the energy using a novel score-based distance that measures discrepancies between samples leveraging the pretrained diffusion score model. By comparing score predictions across multiple noise levels, this metric captures both structural and high-frequency differences that conventional perceptual metrics fail to reflect.

D2PO replaces the static teacher framework with a dynamic reference mechanism that evolves alongside the student policy. Specifically, at each training step, the preference pair is constructed by executing the current policy under two different computational budgets: the losing sample is generated using a fast, sparse timestep schedule, while the winning sample is obtained via a denser, more refined schedule of the same policy. Instead of imitating an external, immutable target, the student is encouraged to align with its own high-quality, dense-schedule approximation. This dynamic preference loop eliminates the fixed error floor inherent in static distillation and implicitly drives the sampler to minimize discretization errors, thereby promoting highly accurate and self-improving sampling trajectories.

Our contributions are summarized as follows:

- We propose D2PO, a preference-based framework for optimizing diffusion samplers, establishing a tractable alignment objective by modeling the deterministic policy as an energy-based surrogate.
- We formulate a novel score-based energy metric derived from the pretrained score network, providing a multi-scale learning signal that captures fine-grained textural and structural details beyond conventional perceptual losses.
- We introduce a dynamic preference mechanism that replaces static teacher supervision with a refinement-based target, enabling continual self-improvement without being bounded by a fixed residual error.
- We comprehensively validate that D2PO learns superior sampling policies, outperforming state-of-the-art distillation-based baselines under various experimental settings.

2 Related Work

2.1 Optimizing diffusion sampling parameters

Since the trajectory of time steps profoundly impacts generation quality under a fixed computational budget, substantial research has focused on finding optimal sampling schedules. Early heuristic approaches, such as EDM [21], employ polynomial spacing to densify steps near the clean data manifold, while Watson *et al.* [52] introduce a dynamic programming framework to search for optimal discrete schedules that maximize log-likelihood. Analytic-DPM [1] improves efficiency by deriving training-free, optimal analytical forms for reverse variances directly from the pretrained score network, while obtaining the corresponding optimal trajectory via dynamic programming [52]. To automate and generalize schedule optimization, AutoDiffusion [25] employs an evolutionary search targeted at minimizing FID, while DDSS [53] optimizes sampler parameters via direct sample-quality feedback such as KID [2].

Another line of work derives analytical error bounds or geometric properties of ODE/SDE trajectories to optimize time discretization. Methods such as those by Chen *et al.* [5, 6], AYS [37], and Xue *et al.* [56] dynamically adjust step sizes based on trajectory curvature or upper bounds of solver errors. More recently, LD3 [46] adopts a relaxed matching objective to learn discretized trajectories through student-teacher regression.

Beyond timestep optimization, recent literature explores tuning other sampling parameters to further accelerate inference. For instance, Galashov *et al.* [13] learn time-dependent CFG weights via a self-consistency objective. Similarly, S4S [12] optimizes solver coefficients at each step using teacher-student matching. Extending this direction, ConsistencySolver [49] employs a learnable high-order solver to dynamically predict optimal integration coefficients.

2.2 Aligning pretrained models with preferences

Driven by the limitations of predefined training objectives, aligning generative models directly with pairwise human or AI preferences has emerged as a dominant paradigm. This approach originated in large language models via Reinforcement Learning from Human Feedback (RLHF) [34], which optimizes policies using a separate reward model. To simplify this multi-stage pipeline, Direct Preference Optimization (DPO) [35] integrates the reward implicitly into the classification loss, enabling stable and direct policy updates. Subsequent self-play frameworks like SPIN [7] further remove the need for preference annotations, generating negatives from the model itself and contrasting them with SFT responses.

Recently, this preference alignment paradigm has been actively adapted to text-to-image diffusion and flow-matching models to enhance visual quality and aesthetic appeal. Standard post-training methods, including DPO-style formulations [26, 48, 57, 61] and online reinforcement learning variants [3, 11, 28], predominantly focus on fine-tuning the foundational weights of the denoiser or velocity networks. While effective, optimizing high-dimensional model parameters

is computationally expensive and risks degrading the quality of outputs. In contrast, D2PO keeps the generative backbone frozen and exclusively optimizes the low-dimensional sampling policy, offering a highly lightweight, orthogonal, and complementary solution to existing weight-tuning approaches.

3 Preliminaries

This section briefly reviews score-based diffusion models and Direct Preference Optimization (DPO), which provide the theoretical foundation for our dynamic sampler optimization framework.

3.1 Diffusion probabilistic models

We consider score-based diffusion models [16,44], which transform a data distribution $p_{\text{real}}(\boldsymbol{x}_0)$ into a tractable prior through a gradual noising process. Specifically, we adopt the Variance Preserving (VP) stochastic differential equation (SDE), which is given by

$$d\boldsymbol{x}_t = -\frac{1}{2}\beta(t)\boldsymbol{x}_t\,dt + \sqrt{\beta(t)}\,d\mathbf{w}_t, \tag{1}$$

where $\mathbf{w}_t$ denotes a standard Wiener process and $t \in [0, T]$. This forward process admits a closed-form marginal:

$$p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}\big(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \mathbf{I}\big)\,, \tag{2}$$

where

$$\alpha_t = \exp\left(-\frac{1}{2}\int_0^t \beta(s)\,ds\right) \qquad \text{and} \qquad \sigma_t^2 = 1 - \alpha_t^2. \tag{3}$$

To construct the reverse process, a neural network $s_\theta(\boldsymbol{x}_t, t)$ is trained to approximate the score function of the marginal distribution, *i.e.*, $s(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ This is achieved via denoising score matching (DSM), which minimizes

$$\mathcal{L}_{\text{DSM}} = \int_0^T \lambda(t)\,\mathbb{E}_{\boldsymbol{x}_0 \sim p_{\text{real}},\, \boldsymbol{x}_t \sim p_t(\cdot|\boldsymbol{x}_0)} \left[\left\| s_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) \right\|_2^2\right] dt, \tag{4}$$

where

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) = -\frac{\boldsymbol{x}_t - \alpha_t \boldsymbol{x}_0}{\sigma_t^2}. \tag{5}$$

Once s_θ is trained, samples are generated by solving the corresponding reverse-SDE from $t = T$ to $t = 0$, which is given by

$$d\boldsymbol{x}_t = \left[-\frac{1}{2}\beta(t)\boldsymbol{x}_t - \beta(t)\,s_\theta(\boldsymbol{x}_t, t)\right] dt + \sqrt{\beta(t)}\,d\bar{\mathbf{w}}_t, \tag{6}$$

where $\bar{\mathbf{w}}_t$ is a reverse-time Wiener process.

This score-based formulation is equivalent to the noise-prediction parameterization $\epsilon_\theta(\boldsymbol{x}_t, t)$ in DDPM [16]. The two representations are related by

$$s_\theta(\boldsymbol{x}_t, t) = -\frac{\epsilon_\theta(\boldsymbol{x}_t, t)}{\sigma_t}. \tag{7}$$

3.2 Direct preference optimization

We build upon Direct Preference Optimization (DPO) [35], a framework for aligning generative policies with preference data. DPO provides a closed-form solution to the KL-regularized reward maximization problem commonly used in RLHF [34], which is defined as

$$\max_{\pi} \mathbb{E}_{\mathbf{x}\sim\pi(\cdot|c)}\big[r(\mathbf{x}, c)\big] - \beta D_{\mathrm{KL}}(\pi(\cdot|c) \,\|\, \pi_{\mathrm{ref}}(\cdot|c)), \tag{8}$$

where π_{ref} is a reference policy and $\beta(>0)$ controls the strength of regularization.

Rather than explicitly learning a reward model $r(\mathbf{x}, c)$ under the context c and performing reinforcement learning, DPO leverages preference pairs $(\mathbf{x}_w, \mathbf{x}_l)$, where $\mathbf{x}_w$ is preferred over $\mathbf{x}_l$. Specifically, by analyzing the optimal solution of Eq. (8), the reward difference between two samples can be expressed via the log-likelihood ratio of the optimal policy relative to the reference policy.This leads to a logistic classification objective on preference pairs as

$$\mathcal{L}_{\mathrm{DPO}}(\phi) = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\Big(\log\frac{\pi_\phi(\mathbf{x}_w|c)}{\pi_{\mathrm{ref}}(\mathbf{x}_w|c)} - \log\frac{\pi_\phi(\mathbf{x}_l|c)}{\pi_{\mathrm{ref}}(\mathbf{x}_l|c)}\Big)\right)\right], \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This formulation eliminates the need for an explicit reward model, directly updating the policy π_ϕ to maximize the relative log-likelihood of preferred samples over unpreferred ones while remaining anchored to the reference policy.

4 D2PO: Dynamic Direct Preference Optimization

4.1 Problem formulation

We aim to optimize the sampling policy of a pretrained diffusion model s_θ by learning a set of sampler parameters $\phi = \{\mathcal{S}, \boldsymbol{\omega}\}$, where $\mathcal{S}$ and $\boldsymbol{\omega}$ denote the timestep schedule and the per-step classifier-free guidance (CFG) weights, respectively. Given a prompt c and an initial noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, our sampler, *i.e.*, ODE solver, deterministically generates an image as

$$\boldsymbol{x}_\phi = \boldsymbol{\Phi}_\phi(\boldsymbol{x}_T, c; \theta), \tag{10}$$

where $\boldsymbol{\Phi}_\phi$ denotes a fixed numerical solver parameterized by ϕ. Although $\boldsymbol{\Phi}_\phi$ is deterministic, it induces a conditional distribution over generated images through the randomness of the initial noise $\boldsymbol{x}_T$:

$$q_\phi(\boldsymbol{x}|c) = \int \delta(\boldsymbol{x} - \boldsymbol{\Phi}_\phi(\boldsymbol{x}_T, c; \theta))\, p(\boldsymbol{x}_T)\, d\boldsymbol{x}_T. \tag{11}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

Our goal is to align $q_\phi(\boldsymbol{x}|c)$ with perceptually preferred outputs using pairwise preference tuples $(c, \boldsymbol{x}_T, \boldsymbol{x}_w, \boldsymbol{x}_l)$, where $\boldsymbol{x}_w$ is preferred over $\boldsymbol{x}_l$. A central challenge is that $q_\phi(\boldsymbol{x}|\boldsymbol{x}_T, c) = \delta(\boldsymbol{x} - \boldsymbol{x}_\phi)$ is degenerate, rendering its likelihood non-differentiable and direct preference optimization ill-posed. We address this by introducing a tractable surrogate formulation.

4.2 Energy-based surrogate for deterministic policies

To apply Direct Preference Optimization (DPO) to our sampler, we replace the $q_\phi(\boldsymbol{x}|\boldsymbol{x}_T, c)$ with a smooth surrogate policy $\pi_\phi(\boldsymbol{x}|c, \boldsymbol{x}_T)$, defined as an Energy-Based Model (EBM):

$$\pi_\phi(\boldsymbol{x}|c, \boldsymbol{x}_T) = \frac{1}{Z(\phi, c, \boldsymbol{x}_T)} \exp(-\alpha E(\boldsymbol{x}; \phi, c, \boldsymbol{x}_T)), \tag{12}$$

where Z is the partition function and α (> 0) a temperature parameter. We define the energy as a distance to the sampler output, which is given by

$$E(\boldsymbol{x}; \phi, c, \boldsymbol{x}_T) \equiv d(\boldsymbol{x}, \boldsymbol{x}_\phi), \tag{13}$$

where $d(\cdot, \cdot)$ is a predefined distance metric.

This surrogate assigns high probability to images close to the sampler's output $\boldsymbol{x}_\phi$ and smoothly decays as the proximity decreases. Such functional relaxations are commonly employed to bypass the non-differentiability of objective functions for optimization. For example, score-based models apply Gaussian perturbations—mathematically forming EBMs with ℓ^2 energy—to resolve undefined gradients. Soft Actor-Critic [14] and the Gumbel–Softmax [19] use probabilistic relaxations on discrete policies and operations to enable backpropagation. Our surrogate plays a similar role for generative optimization: it replaces the non-differentiable Dirac delta function with a smooth landscape whose mode coincides with $\boldsymbol{x}_\phi$, thereby binding updates to the true generative process while keeping the objective differentiable.

Importantly, when computing the log-probability ratio between two candidates sharing the same context $(c, \mathbf{x}_T)$, the identical partition function $Z(\phi, c, \mathbf{x}_T)$ cancels out, allowing the ratio to simplify exactly as follows:

$$\log \frac{\pi_\phi(\boldsymbol{x}_w|c, \boldsymbol{x}_T)}{\pi_\phi(\boldsymbol{x}_l|c, \boldsymbol{x}_T)} = -\alpha \left(E(\boldsymbol{x}_w; \phi, c, \boldsymbol{x}_T) - E(\boldsymbol{x}_l; \phi, c, \boldsymbol{x}_T)\right). \tag{14}$$

By applying this to both the student sampler ϕ and the reference sampler ϕ_{ref} and substituting this expression into the DPO objective (Eq. (9)), we derive the final D2PO objective:

$$\mathcal{L}_{\text{D2PO}}(\phi) = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\beta \Delta_\phi(\boldsymbol{x}_w, \boldsymbol{x}_l)\right)\right], \tag{15}$$

where

$$\Delta_\phi(\boldsymbol{x}_w, \boldsymbol{x}_l) = \big(d(\boldsymbol{x}_w, \boldsymbol{x}_{\phi_{\text{ref}}}) - d(\boldsymbol{x}_w, \boldsymbol{x}_\phi)\big) - \big(d(\boldsymbol{x}_l, \boldsymbol{x}_{\phi_{\text{ref}}}) - d(\boldsymbol{x}_l, \boldsymbol{x}_\phi)\big),$$

and ϕ_{ref} denotes the reference policy. The temperature parameter α is implicitly absorbed into the scaling factor β for simplicity.

4.3 Score-based distance

The effectiveness of D2PO depends on the choice of the distance function $d(\cdot, \cdot)$ used in the energy definition of the surrogate policy π_ϕ. A naïve choice would adopt a predefined metric such as ℓ_2 or LPIPS, or train a separate network to approximate the energy. However, such choices fail to exploit the rich representations already encoded within the pretrained diffusion model.

Score-induced energy. Our key insight is to derive the energy directly from the pretrained score network $s_\theta(\boldsymbol{x}_t, t)$. Recall that s_θ approximates the data score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ at noise level t. Since the score characterizes the geometry of the data distribution, it naturally quantifies sample likelihood.

A sample that lies on the true data manifold should be locally consistent with this score field. Conversely, a sample out of the data distribution fails to align with the score trajectory. We therefore define the ideal score-induced energy of a sample as the degree of its misalignment with the learned score geometry as follows:

$$E(\boldsymbol{x}; \phi) \equiv \int_0^T w(t) \, \|s_\theta(\boldsymbol{x}_t, t) - s_\theta(\boldsymbol{x}_{\phi,t}, t)\|_2^2 \, dt, \tag{16}$$

where $w(t)$ is a weighting function over noise levels, and $\boldsymbol{x}_t$ and $\boldsymbol{x}_{\phi,t}$ denote the perturbed versions of $\boldsymbol{x}$ and $\boldsymbol{x}_\phi$ at noise level t.

We evaluate the score discrepancy over the perturbed data distributions rather than the clean data distribution ($t = 0$) for both theoretical and practical reasons. In standard score-based generative modeling [44, 51, 60], the clean data score is unavailable and pretrained diffusion models do not directly learn this clean score. Instead, they are trained to approximate the scores of perturbed distributions p_t across a continuous spectrum of noise levels. These noisy score fields encode the multi-scale geometry of the data manifold, capturing coarse semantic structures at large t and fine-grained details at small t. Leveraging these perturbed distributions is therefore tractable and consistent with the objective of the pretrained diffusion model.

Noise-prediction distance. To transform the ideal energy formulation in Eq. (16) into a practical optimization objective, we reframe the score discrepancy via the noise-prediction error as follows:

$$d_\theta(\mathbf{x}, \mathbf{x}_\phi; t) = \|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon_\theta(\mathbf{x}_{\phi,t}, t)\|_2^2 \tag{17}$$

which leverages the implicit relation, $s_\theta = -\epsilon_\theta / \sigma_t$. Directly substituting this noise-prediction distance for the score-based distance in Eq. (16) under a uniform weighting, however, introduces a scale mismatch across different noise levels. Specifically, because the score discrepancy equals the noise-prediction distance up to a scaling factor of σ_t^{-2}, the integrand tends to diverge numerically as $t \to 0$, causing the low-noise terms to dominate the overall energy. To resolve this imbalance and stabilize the optimization, we follow the established practice in DDPM [16] by adopting the weighting function $w(t) = \sigma_t^2$, which cancels

the σ_t^{-2} factor; the score-induced energy in Eq. (16) reduces to the total noise-prediction distance over the continuous trajectory, which is given by

$$E(\mathbf{x};\phi) = \int_0^T d_\theta(\mathbf{x}, \mathbf{x}_\phi; t)\, dt. \tag{18}$$

As evaluating this continuous integral is computationally expensive, we approximate it in practice via Monte Carlo sampling.

Practical D2PO objective. Substituting the weighted score-based distance for the surrogate policy π_ϕ and applying the DPO objective between the student policy π_ϕ and the reference policy $\pi_{\phi_{\text{ref}}}$, we obtain the following objective:

$$\mathcal{L}_{\text{D2PO}}(\phi) = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\Big(\mathbb{E}_{t\sim\mathcal{U}(0,T)}\Big[\beta\Delta_\phi(\boldsymbol{x}_w, \boldsymbol{x}_l; t)\Big]\Big)\right], \tag{19}$$

where

$$\Delta_\phi(\boldsymbol{x}_w, \boldsymbol{x}_l; t) = \big(d_\theta(\boldsymbol{x}_w, \boldsymbol{x}_{\phi_{\text{ref}}}; t) - d_\theta(\boldsymbol{x}_w, \boldsymbol{x}_\phi; t)\big) - \big(d_\theta(\boldsymbol{x}_l, \boldsymbol{x}_{\phi_{\text{ref}}}; t) - d_\theta(\boldsymbol{x}_l, \boldsymbol{x}_\phi; t)\big).$$

A key benefit of the proposed noise-prediction distance lies in its evaluation over multi-level noise. The score field captures coarse semantic structure at large noise levels t and fine-grained detail at small t, and the weighting $w(t) = \sigma_t^2$ aggregates these scales into a single well-conditioned signal over t. This yields a significantly richer preference signal compared to perceptual metrics such as LPIPS.

4.4 Dynamic preference

To understand D2PO and its dynamic optimization mechanism, it is essential to identify three key components: the *student sampler*, the *reference sampler*, and the *winning sampler*, which are parameterized by ϕ, ϕ_{ref}, and ϕ', respectively.

The student sampler, governed by the target parameters $\phi = \{\mathcal{S}, \boldsymbol{\omega}\}$, represents the policy we aim to optimize. To establish a preference comparison for DPO training, we utilize its output $\boldsymbol{x}_\phi$ as a baseline rather than relying on an external target. Specifically, the losing sample $\boldsymbol{x}_l$ is synthesized by applying a degradation operator $\mathcal{G}$ (e.g., a low-pass filter) to the student output, expressed as $\boldsymbol{x}_l = \mathcal{G}(\text{sg}[\boldsymbol{x}_\phi])$, where $\text{sg}[\cdot]$ denotes the stop-gradient operator.

Unlike standard DPO which employs a static reference model, D2PO dynamically updates the reference parameters $\phi_{\text{ref}} = \{\mathcal{S}_{\text{ref}}, \boldsymbol{\omega}_{\text{ref}}\}$. Inspired by SPIN [7], the reference timestep schedule $\mathcal{S}_{\text{ref}}$ is synchronized by copying the student schedule $\mathcal{S}$ at the end of each epoch. Meanwhile, the reference CFG weights $\boldsymbol{\omega}_{\text{ref}}$ are adjusted at each training step via an Exponential Moving Average (EMA) with a momentum parameter λ, i.e., $\boldsymbol{\omega}_{\text{ref}} \leftarrow \lambda\boldsymbol{\omega}_{\text{ref}} + (1-\lambda)\boldsymbol{\omega}$.

Designing a dynamic winning sampler is a core contribution of D2PO. Instead of introducing a pre-computed, fixed teacher, we formulate a dynamic teacher sampler whose implied distribution, $\pi_{\phi'}$, is generated relative to the current student parameters ϕ at each training step. For instance, if the student schedule $\mathcal{S}$ dictates a coarse trajectory with N timesteps, we construct

the dynamic teacher's schedule $\mathcal{S}'$ by refining $\mathcal{S}$ with additional intermediate timesteps, yielding a denser $2N$-step trajectory (e.g., via linear interpolation). The dynamic teacher sampler, parameterized by ϕ', is then induced by the same numerical solver operating under this finer-grained schedule $\mathcal{S}'$.

This dynamic framework provides a more robust learning signal than a static teacher policy (π^{fix}). By design, the dynamic teacher represents a higher-fidelity trajectory derived from the student's current parameters. Consequently, the D2PO loss penalizes the discrepancy between the student's coarse numerical path and this refined counterpart. This formulation encourages the student sampler to yield a trajectory that remains consistent under step-size refinement, which is achieved when the discrete path closely approximates the true continuous-time trajectory. Ultimately, rather than tracking an arbitrary external target, the student effectively learns to minimize its own discretization error.

4.5 Theoretical analysis

We provide a theoretical justification for the efficacy of the dynamic teacher mechanism in D2PO, thereby reducing discretization error.

Setup Let π denote the true continuous-time policy representing the target distribution. We define π_ϕ as the student policy induced by the parameters ϕ under a discrete numerical schedule with N timesteps. The dynamic teacher corresponds to a refined policy $\pi_{\phi'}$ evaluated on a finer discretization schedule (e.g., $2N$ timesteps). To quantify discrepancies between policies, we employ a metric $\rho(\cdot, \cdot)$ that satisfies the triangle inequality. The true error of the student policy relative to the continuous-time target is defined as

$$\epsilon_\phi^{\text{true}} = \rho(\pi_\phi, \pi). \tag{20}$$

Dynamic teacher The dynamic DPO objective minimizes the discrepancy between the student and its refined counterpart:

$$\mathcal{L}_{\text{dyn}} = \rho(\pi_\phi, \pi_{\phi'}). \tag{21}$$

Assuming the underlying numerical solver exhibits a convergence order of $k > 0$ [45], the dynamic teacher $\pi_{\phi'}$ constructed via a $2\times$ refinement yields a reduced true error relative to the continuous-time target, which is given by

$$\epsilon_{\phi'}^{\text{true}} = \rho(\pi_{\phi'}, \pi) \approx \frac{1}{2^k}\, \epsilon_\phi^{\text{true}}. \tag{22}$$

By applying the triangle inequality, $\rho(\pi_\phi, \pi) \leq \rho(\pi_\phi, \pi_{\phi'}) + \rho(\pi_{\phi'}, \pi)$, we establish a lower bound on the dynamic loss:

$$\mathcal{L}_{\text{dyn}} \geq \left|\epsilon_\phi^{\text{true}} - \epsilon_{\phi'}^{\text{true}}\right| \approx \left(1 - \frac{1}{2^k}\right) \epsilon_\phi^{\text{true}}. \tag{23}$$

Consequently, $\mathcal{L}_{\text{dyn}}$ serves as a non-trivial surrogate that upper-bounds (and scales proportionally with) the student's true error. Minimizing $\mathcal{L}_{\text{dyn}}$ enforces trajectory consistency across different discretization granularities, thereby driving a systematic reduction in discretization error. Since $\epsilon_{\phi'}^{\text{true}} = O(2^{-k})$, the residual error of the teacher vanishes progressively under refinement, aligning $\mathcal{L}_{\text{dyn}}$ more closely with the true optimization objective as training proceeds.

Fixed teacher The conventional fixed-teacher objective minimizes

$$\mathcal{L}_{\text{fix}} = \rho(\pi_\phi, \pi^{\text{fix}}), \tag{24}$$

where π^{fix} denotes a static teacher policy that is independent of the student parameters ϕ. Let the intrinsic error of this fixed teacher be

$$\epsilon^{\text{fix}} = \rho(\pi^{\text{fix}}, \pi), \tag{25}$$

which remains constant with respect to ϕ. Applying the triangle inequality yields the following bounds on the empirical loss:

$$\left|\epsilon_\phi^{\text{true}} - \epsilon^{\text{fix}}\right| \leq \mathcal{L}_{\text{fix}} \leq \epsilon_\phi^{\text{true}} + \epsilon^{\text{fix}}. \tag{26}$$

Even if $\mathcal{L}_{\text{fix}}$ is optimized to its global minimum ($\mathcal{L}_{\text{fix}} = 0$), the resulting student policy is bounded by $\epsilon_\phi^{\text{true}} = \epsilon^{\text{fix}}$. Therefore, minimizing $\mathcal{L}_{\text{fix}}$ cannot reduce the true error below this asymptotic error floor. D2PO bypasses this performance bottleneck because the dynamic teacher evolves alongside the student, preventing the optimization from stagnation at a fixed residual error floor.

5 Experiment

5.1 Experimental setup

We comprehensively evaluate D2PO across multiple generation tasks, architectures, and datasets. For text-to-image synthesis, we use the pre-trained Stable Diffusion v1.5 model [36] using prompts from the COCO [27] dataset. To demonstrate the generalizability of our approach, we apply this exact same sampler optimization to ImageNet (256×256) generation in the latent space, as well as to the sampling process of the flow-matching-based InstaFlow model. Crucially, we keep all pre-trained model parameters strictly frozen; our method exclusively optimizes the sampling policy, consisting of the continuously parameterized timestep schedule $\mathcal{S}$ and the per-step CFG weights $\boldsymbol{\omega}$ for several advanced ODE solvers, including iPNDM [64], Uni_PC [66] and DPM_solver++ [32]. We compare D2PO against state-of-the-art discretization methods, including DMN [56], GITS [5], and LD3 [46]. For evaluation, we measure distributional fidelity using FID [15], and further assess text-to-image perceptual quality using HPSv2 [54] and Aesthetic scores [40][1].

[1]Baselines for text-to-image tasks are re-evaluated on newly generated samples since HPSv2 and Aesthetic scores are omitted in the original papers. For ImageNet, we copy the FID values reported in their respective papers.

Table 1: Quantitative comparison of sampler optimization methods on text-to-image synthesis using Stable Diffusion v1.5 across different ODE solvers.

		iPNDM			**UniPC**			**DPM Solver++**		
Steps	Method	HPS ↑	Aesthetic ↑	FID ↓	HPS ↑	Aesthetic ↑	FID ↓	HPS ↑	Aesthetic ↑	FID ↓
4	DMN [56]	0.2030	5.0936	21.39	0.2030	5.1084	22.03	0.1979	5.0978	24.33
	GITS [5]	0.2128	5.1413	18.12	0.2109	5.1592	20.14	0.2096	5.1519	19.86
	LD3 [46]	0.2191	5.1756	17.60	0.2180	5.1755	18.34	0.2191	5.1736	17.46
	D2PO	**0.2237**	**5.2024**	**15.69**	**0.2185**	**5.1761**	**16.97**	**0.2216**	**5.1854**	**16.84**
5	DMN [56]	0.2146	5.1514	17.35	0.2175	5.1732	16.99	0.2108	5.1549	19.17
	GITS [5]	0.2274	5.1926	14.07	0.2268	5.1928	15.46	0.2260	5.2058	15.29
	LD3 [46]	0.2346	5.2463	13.59	0.2355	**5.2591**	**13.89**	0.2344	5.2378	**13.27**
	D2PO	**0.2385**	**5.2701**	**13.38**	**0.2369**	5.2776	14.47	**0.2374**	**5.2740**	14.16
6	DMN [56]	0.2283	5.2107	13.66	0.2322	5.2371	13.74	0.2267	5.2188	14.51
	GITS [5]	0.2382	5.2407	**12.33**	0.2404	5.2461	**12.38**	0.2397	5.2610	**12.41**
	LD3 [46]	0.2375	5.2565	13.10	0.2402	5.2689	12.90	0.2386	5.2445	12.62
	D2PO	**0.2482**	**5.3309**	13.54	**0.2441**	**5.3192**	14.38	**0.2458**	**5.3237**	14.00
7	DMN [56]	0.2428	5.2778	**11.89**	0.2474	5.3011	**12.12**	0.2399	5.2734	13.02
	GITS [5]	0.2394	5.2368	12.16	0.2379	5.2065	12.91	0.2355	5.2174	13.16
	LD3 [46]	0.2434	5.2763	12.41	0.2421	5.2660	12.77	0.2455	5.2795	**12.09**
	D2PO	**0.2513**	**5.3257**	12.71	**0.2499**	**5.3472**	13.61	**0.2502**	**5.3458**	13.52

5.2 Main results

Our quantitative results on text-to-image synthesis using Stable Diffusion v1.5 are presented in Tab. 1, where we apply D2PO to three representative ODE solvers—iPNDM, UniPC, and DPM Solver++. For evaluation, we generate 30k samples using text prompts from COCO dataset [27], following the standard [12, 46]. Overall, the results demonstrate that D2PO achieves superior performance compared to existing baselines, asserting its effectiveness and robustness across multiple advanced solvers.

At low number of steps (4 and 5), D2PO functions as a superior error corrector. In this regime, D2PO achieves state-of-the-art perceptual quality—showing the highest Aesthetic scores across all solvers and the highest HPS scores for iPNDM. Furthermore, when using 4 steps, D2PO surpasses all baselines in FID across all three solvers. This directly validates of our hypothesis. Regression-based methods sacrifice high-frequency details, in part because their loss metrics, *e.g.*, LPIPS, fail to capture these errors. Our novel score-based distance metric captures both structural and textural discrepancies by measuring score differences at various noise levels. This high sensitivity to fine-grained texture loss allows D2PO to correct the foundational flaws of the baseline, simultaneously improving both quality and fidelity.

At higher number of steps (6 and 7), as the baseline's severe discretization error is reduced, the expected quality-diversity trade-off [4, 17, 24] emerges, and D2PO's behavior shifts to its primary goal of preference alignment. While D2PO maintains its significant lead in perceptual scores (HPS/Aesthetic), its FID score becomes comparable or slightly higher than the baselines. This shift is the expected signature of successful alignment. D2PO's objective function,

Table 2: Quantitative comparison on ImageNet-256 (latent space) using the 3rd-order (3M) iPNDM solver. Baseline results are reported from the LD3 [46] paper.

	Steps			
Method	4	5	6	7
Uniform	13.86	7.80	6.03	5.35
GITS [5]	56.00	43.56	19.33	10.33
DMN [56]	10.15	7.33	7.25	7.40
LD3 [46]	9.19	6.03	5.09	**4.68**
D2PO	**7.28**	**5.48**	**4.80**	4.70

Table 3: Quantitative comparison on the InstaFlow [30] model using the prompts from the COCO dataset. Higher HPS and Aesthetic scores (↑) are better, while lower FID scores (↓) are better.

Steps	Method	HPS ↑	Aesthetic ↑	FID ↓
2	Time Uniform	0.1865	5.0060	44.27
	LD3 [46]	0.1708	4.6270	63.55
	D2PO	**0.1872**	**5.0924**	**40.68**
4	Time Uniform	0.2189	5.1222	16.85
	LD3 [46]	0.2086	5.0712	23.21
	D2PO	**0.2197**	**5.1415**	**15.50**
6	Time Uniform	0.2317	5.1847	13.68
	LD3 [46]	0.2299	5.1740	15.60
	D2PO	**0.2342**	**5.1855**	**12.70**

Table 4: Ablation study of D2PO on text-to-image synthesis using Stable Diffusion v1.5 and the iPNDM solver using prompts from the COCO dataset. We report performance at # steps=4 and 5 after removing or altering key components: dynamic preference, score-based energy, and the reference timestep update strategy.

Steps	Method	Aesthetic ↑	FID ↓
4	D2PO (Full)	**5.2024**	**15.69**
	w/o dynamic preference	5.1810	16.91
	w/o score-based energy	5.1796	17.88
	w/ EMA reference timestep	5.1937	15.75
5	D2PO (Full)	**5.2701**	**13.38**
	w/o dynamic preference	5.2615	13.70
	w/o score-based energy	5.2630	14.59
	w/ EMA reference timestep	5.2639	13.42

which aligns with its dynamic, higher-fidelity $2N$ teacher, optimizes the sampler towards a distribution that maximizes perceptual quality. This distribution is distinct from the average of the real data distribution, which FID measures. This result demonstrates that D2PO is not failing; rather, it is successfully harnessing the full potential of the model by optimizing for its intended perceptual objective, which is to reduce its own discretization error.

5.3 Generalization across domains and architectures

To evaluate the robustness of our learned policy, we tested D2PO beyond standard text-to-image tasks. As shown in Tab. 2, D2PO consistently outperforms baseline methods on ImageNet-256 (latent space) using the 3rd-order iPNDM solver. Furthermore, Tab. 3 illustrates D2PO's successful application to InstaFlow, a flow-matching model on the COCO dataset.

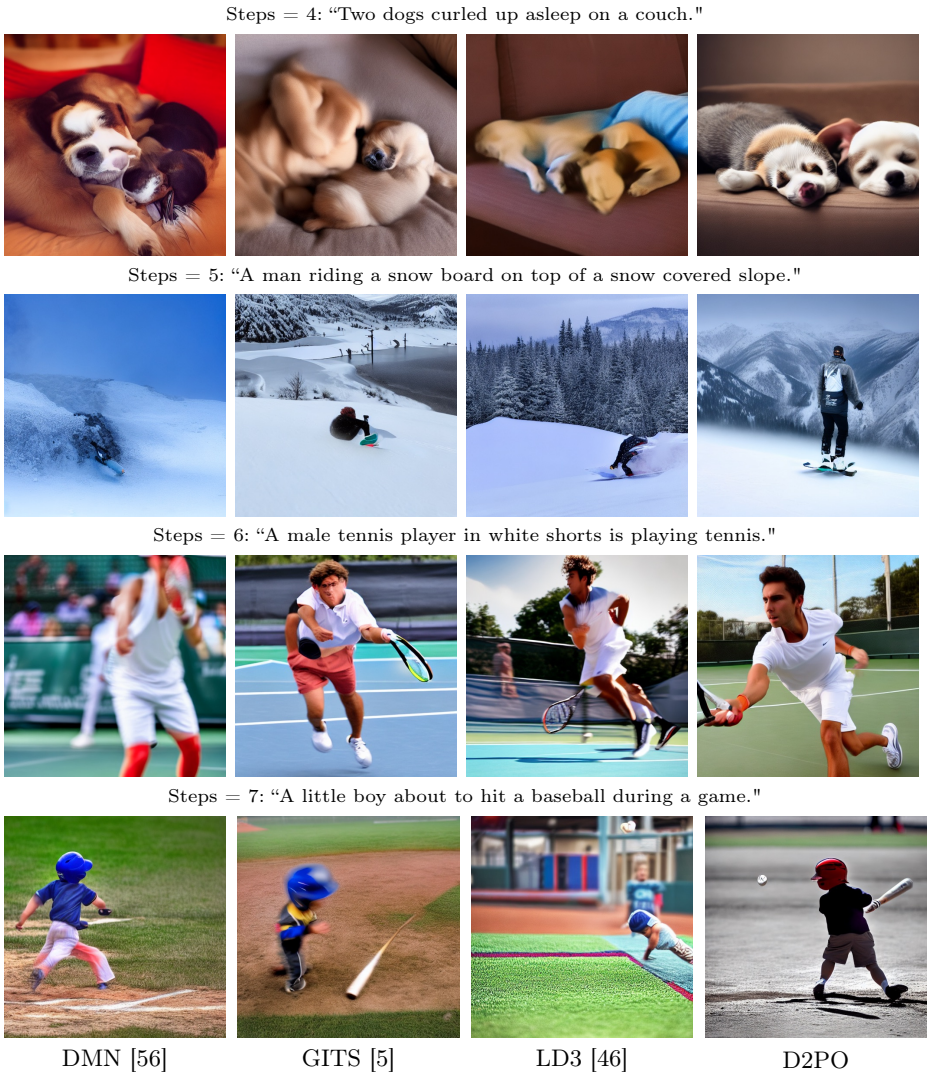

Fig. 2: Qualitative comparison of various discretization methods on Stable Diffusion v1.5 using iPNDM across different number of time steps (4 to 7). D2PO consistently produces sharper details and fewer artifacts compared to baselines.

5.4 Qualitative results

Fig. 2 visualize the qualitative results on Stable Diffusion v1.5 (iPNDM) for all steps reported in Tab. 1. Across all cases, D2PO consistently produces images with higher perceptual quality, sharper details, and fewer artifacts compared to baselines. This visual evidence directly supports our quantitative findings. For instance, baseline methods like LD3 and DMN frequently exhibit significant

blurriness, water-color artifacts, or loss of fine-grained detail. In more complex scenes (Steps=7), D2PO generates a coherent and detailed image, while other methods suffer from severe structural distortion or artifacts. More qualitative results on different settings are provided in the supplementary material.

5.5 Ablation study

We ablate D2PO's key components in Tab. 4. First, the w/o dynamic preference variant replaces our dynamic $2N$-step teacher with a fixed $N+1$-step teacher (as in LD3 [46]). This degrades performance, validating that our dynamic mechanism provides a stronger, more consistent learning signal than regression to static teacher. Second, the w/o score-based energy variant, which instead use LPIPS, causes the most significant drop in fidelity, confirming our metric's necessity for capturing fine-grained discretization errors. Finally, the w/ EMA reference timestep variant replaces our epoch-wise copy strategy for the reference time step schedule with an EMA update. Its sub-optimal performance highlights that while EMA suits continuous parameters (like CFG weights), direct epoch-wise copying provides a more stable anchor for optimizing discrete time steps.

6 Conclusion

We identified a critical performance bottleneck in dominant student-teacher regression frameworks for optimizing diffusion samplers. We demonstrated that as the teacher-student NFE gap increases, standard regression losses force the low-NFE student to sacrifice high-frequency texture fidelity, leading to degraded perceptual quality. To address this, we proposed D2PO, a novel framework that reframes sampler optimization as a preference-based alignment task. We introduced a novel score-based energy function that leverages the score model itself to capture the fine-grained textural and structural errors that standard metrics miss, and a dynamic preference mechanism that creates a self-improving loop, where the student policy is aligned with a dynamically refined, higher-fidelity version of itself. This dynamic teacher provides a stronger, theoretically-grounded learning signal that forces the student to minimize its own discretization error, rather than converging to a suboptimal fixed teacher. Our extensive experiments across multiple solvers demonstrated that D2PO successfully aligns diffusion samplers with true perceptual quality, effectively solving the existing bottleneck of static teacher regression.

Acknowledgements. This work was partly supported by the Samsung Electronics Co., Ltd. (IO250418-12669-01). It was also partly supported by the NRF grant [RS-2022-NR070855] and the IITP grants [RS-2025-25442338; RS-2026-25526850; No.RS-2021-II211343; No.RS-2020-II201336] funded by the Korea government (MSIT).

References

1. Bao, F., Li, C., Zhu, J., Zhang, B.: Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. In: ICLR (2022)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: ICLR (2018)
3. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. In: ICLR (2023)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
5. Chen, D., Zhou, Z., Wang, C., Shen, C., Lyu, S.: On the trajectory regularity of ODE-based diffusion sampling. In: ICML (2024)
6. Chen, Y., He, F., Fu, S., Tian, X., Tao, D.: Adaptive time-stepping schedules for diffusion models. In: The 40th Conference on Uncertainty in Artificial Intelligence (2024)
7. Chen, Z., Deng, Y., Yuan, H., Ji, K., Gu, Q.: Self-play fine-tuning converts weak language models to strong language models. In: ICML. PMLR (2024)
8. Choi, J., Kang, J., Han, B.: Enhanced diffusion sampling via extrapolation with multiple ode solutions. In: ICLR (2025)
9. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS (2021)
10. Dockhorn, T., Vahdat, A., Kreis, K.: GENIE: Higher-order denoising diffusion solvers. In: NeurIPS (2022)
11. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Reinforcement learning for fine-tuning text-to-image diffusion models. In: NeurIPS (2024)
12. Frankel, E., Chen, S., Li, J., Koh, P.W., Ratliff, L.J., Oh, S.: S4s: Solving for a fast diffusion model solver. In: ICML (2025)
13. Galashov, A., Pokle, A., Doucet, A., Gretton, A., Delbracio, M., De Bortoli, V.: Learn to guide your diffusion model. arXiv preprint arXiv:2510.00815 (2025)
14. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: ICML (2018)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: arXiv preprint arXiv:2207.12598 (2022)
18. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022)
19. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-Softmax. In: ICLR (2017)
20. Kang, J., Choi, J., Choi, S., Han, B.: Observation-guided diffusion probabilistic models. In: CVPR (2024)
21. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
22. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In: ICLR (2024)

23. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: NeurIPS (2021)
24. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. NeurIPS (2018)
25. Li, L., Li, H., Zheng, X., Wu, J., Xiao, X., Wang, R., Zheng, M., Pan, X., Chao, F., Ji, R.: AutoDiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In: ICCV (2023)
26. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In: CVPR (2025)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (2014)
28. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. In: NeurIPS (2025)
29. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: ICLR (2022)
30. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In: ICLR (2024)
31. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In: NeurIPS (2022)
32. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. In: ICLR (2023)
33. Ma, X., Fang, G., Bi Mi, M., Wang, X.: Learning-to-cache: Accelerating diffusion transformer via layer caching. NeurIPS (2024)
34. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NeurIPS (2022)
35. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: NeurIPS (2024)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
37. Sabour, A., Fidler, S., Kreis, K.: Align your steps: Optimizing sampling schedules in diffusion models. ICML (2024)
38. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR (2022)
39. Salimans, T., Mensink, T., Heek, J., Hoogeboom, E.: Multistep distillation of diffusion models via moment matching. In: NeurIPS (2024)
40. Schuhmann, C.: Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/` (2022), accessed: 2023-11-10
41. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-A-Video: Text-to-video generation without text-video data. In: ICLR (2022)
42. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
43. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML (2023)
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)

45. Süli, E., Mayers, D.: An Introduction to Numerical Analysis. Cambridge University Press, 1 edn. (2003)
46. Tong, V., Trung-Dung, H., Liu, A., Broeck, G.V.d., Niepert, M.: Learning to discretize denoising diffusion ODEs. In: ICLR (2025)
47. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. In: NeurIPS (2021)
48. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: CVPR (2024)
49. Wang, F.Y., Zhou, H., Yuan, L., Woo, S., Gong, B., Han, B., Yang, M.H., Zhang, H., Zhu, Y., Liu, T., Zhao, L.: Image diffusion preview with consistency solver. In: CVPR (2026)
50. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: ModelScope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
51. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
52. Watson, D., Chan, W., Ho, J., Norouzi, M.: Learning fast samplers for diffusion models by differentiating through sample quality. In: ICLR (2022)
53. Watson, D., Chan, W., Ho, J., Norouzi, M.: Learning fast samplers for diffusion models by differentiating through sample quality. In: ICLR (2022)
54. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
55. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion GANs. In: ICLR (2022)
56. Xue, S., Liu, Z., Chen, F., Zhang, S., Hu, T., Xie, E., Li, Z.: Accelerating diffusion sampling with optimized time steps. In: CVPR (2024)
57. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: CVPR (2024)
58. Ye, H., Yuan, J., Xia, R., Yan, X., Chen, T., Yan, J., Shi, B., Zhang, B.: Training-free adaptive diffusion with bounded difference approximation strategy. NeurIPS (2024)
59. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T.: Improved distribution matching distillation for fast image synthesis. In: NeurIPS (2024)
60. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: CVPR (2024)
61. Yuan, H., Chen, Z., Ji, K., Gu, Q.: Self-play fine-tuning of diffusion models for text-to-image generation. NeurIPS (2024)
62. Zhang, G., Kenta, N., Kleijn, W.B.: Lookahead diffusion probabilistic models for refining mean estimation. In: CVPR (2023)
63. Zhang, G., Kenta, N., Kleijn, W.B.: On accelerating diffusion-based sampling process via improved integration approximation. In: ICLR (2024)
64. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: ICLR (2023)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
66. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In: NeurIPS (2023)

67. Zheng, J., Hu, M., Fan, Z., Wang, C., Ding, C., Tao, D., Cham, T.J.: Trajectory consistency distillation. arXiv preprint arXiv:2402.19159 (2024)
68. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: MagicVideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
69. Zhou, M., Zheng, H., Gu, Y., Wang, Z., Huang, H.: Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In: ICLR (2025)
70. Zhou, M., Zheng, H., Wang, Z., Yin, M., Huang, H.: Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In: ICML (2024)