UNDERSTANDING SAMPLER STOCHASTICITY IN TRAINING DIFFUSION MODELS FOR RLHF

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) is increasingly used to fine-tune diffusion models, but a key challenge arises from the mismatch between stochastic samplers used during training and deterministic samplers used during inference. In practice, models are fine-tuned using stochastic SDE samplers to encourage exploration, while inference typically relies on deterministic ODE samplers for efficiency and stability. This discrepancy induces a **reward gap**, raising concerns about whether high-quality outputs can be expected during inference. In this paper, we theoretically characterize this reward gap and provide non-vacuous bounds for general diffusion models, along with sharper convergence rates for Variance Exploding (VE) and Variance Preserving (VP) Gaussian models. Methodologically, we adopt the generalized DDIM (gDDIM) framework to support arbitrarily high levels of stochasticity, preserving data marginals throughout. Empirically, our findings through large-scale experiments on text-to-image models using DDPO and MixGRPO validate that reward gaps consistently narrow over training, and ODE sampling quality improves when models are updated using higher-stochasticity SDE training.

1 INTRODUCTION

Diffusion models (e.g., Stable Diffusion (Rombach et al., 2022), SDXL (Podell et al., 2024), FLUX (Black Forest Labs, 2024)) have shown strong performance in text-to-image (T2I) tasks, and have also been extended beyond images to video (Ho et al., 2022) and audio (Liu et al., 2023). To meet downstream objectives such as aesthetics, safety, and alignment, it is essential to post-train with RLHF (Ouyang et al., 2022) for preference-driven improvements, often with a KL-regularization term to preserve performance on pretrained tasks (Schulman et al., 2017). Widely used RLHF algorithms include DDPO (Black et al., 2024) and GRPO (Shao et al., 2024) variants (FlowGRPO (Liu et al., 2025), DanceGRPO (Xue et al., 2025), MixGRPO (Li et al., 2025)). DDPO directly optimizes human-preference rewards by casting the denoising process as a Markov Decision Process (MDP); GRPO variants use group-relative advantages. See (Winata et al., 2025) for a broader review of successful RLHF algorithms for generative models.

Despite strong progress in alignment, RLHF training often exhibits unstable trajectories, long inference times, and vulnerability to reward hacking (Skalse et al., 2022). Using multiple rewards can mitigate the latter (Lee et al., 2025), but we also need efficient, robust samplers to produce stable, high-quality fine-tuned models. A classical scheme is DDPM (Ho et al., 2020), a discretization of the score-based backward SDE (Risken, 1996; Song et al., 2021b), which preserves data marginals but is time-consuming. In contrast, the deterministic DDIM sampler (Song et al., 2021a) follows the probability-flow ODE, enabling marginal-preserving sampling with fewer steps. DDPO typically uses DDIM with constant stochasticity to generate training samples; MixGRPO mixes SDE and ODE steps, varying stochasticity across the denoising horizon. However, although stochasticity is valuable for generating diverse training data, fine-tuned models often use DDIM or higher-order ODE solvers (Lu et al., 2022; 2025) for fast, stable inference. This naturally raises the question:

Why can we guarantee good sampling quality when inference uses a different noise level than the one used during RLHF training?

Figure 1: ODE (below) image generation preserves prompt instructions with better quality on details compared to SDE (above) image generation under large stochasticity ($\eta = 1.2$).
Prompts (from left to right): "A vintage writing desk with an open journal and a flickering candle.", "A macaque soaking in a steaming hot spring, surrounded by falling snow.", "A wicker rocking chair on a wrap-around porch during golden hour.", "A sleek sports car drifting on a mountain highway during golden hour."

In this paper, we address it from three perspectives. Theoretically, we derive non-vacuous bounds on the reward gap between a generally SDE-fine-tuned model and its ODE-sampling counterpart using Gronwall's inequality. Specifically, for Variance Exploding (VE) and Variance Preserving (VP) Gaussian models, the gap shrinks at sharp rates $O(1/T)$ and $O(e^{-T^2})$, where T is the denoising time horizon. These results justify the practice of ODE inference for SDE-fine-tuned models and provide insights into more randomized exploration beyond $\eta = 1.0$, the noise level for continuous score-based backward dynamics. Methodologically, we adopt the generalized DDIM (gDDIM) (Zhang et al., 2023) framework for arbitrary stochasticity levels. Rather than linear variance interpolations between DDIM and DDPM, our implementation preserves marginals for all $\eta \geq 0$ and, more importantly, supports $\eta > 1$ in a principled way, enabling broader RLHF exploration beyond standard DDPM noise levels. Empirically, we evaluate large-scale T2I models and RLHF algorithms to quantify the reward gap across multiple reward functions. The results show that (i) reward gaps decrease as training quality improves, (ii) moderate-to-high training stochasticity (e.g., $\eta = 1.2$) yields superior in- and out-of-domain performance, and (iii) ODE inference remains stable and often outperforms SDE inference under small denoising step budgets.

- **Theory**: Sharp bounds for VE/VP models; non-vacuous bounds for general distributions.
- **Methodology**: gDDIM-based noise scheduling that preserves marginals for $\eta \geq 0$ and supports $\eta > 1$ in a principled way.
- **Experiments**: Comprehensive evaluations across rewards and RLHF algorithms, validating bounded reward gaps and improved ODE inference with higher training stochasticity.

Related Literature. Higher-order ODE solvers such as RX-DPM (Choi et al., 2025), DEIS (Zhang & Chen, 2022) approximate probability-flow. Recent analyses give broader convergence guarantees (Huang et al., 2025) and better representability (Chen et al., 2023) for score-based diffusion. In parallel, RL-based fine-tuning leverages stochastic sampling: DRaFT differentiates through noisy trajectories (Clark et al., 2023), Score-as-Action casts fine-tuning as stochastic control (Zhao et al., 2025), and Adjoint Matching enforces optimal memoryless noise schedules (Domingo-Enrich et al., 2024). Large-scale studies combine multiple rewards (Zhang et al., 2024); SEPO, D3PO, and ImageReFL boosts alignment (Zekri & Boullé, 2025; Yang et al., 2024; Sorokin et al., 2025). Finally, (Liang et al., 2025) gives discretization error bound and (Wu et al., 2024) discusses guidance (Ho & Salimans, 2021) for Gaussian Mixture priors.

Organization of the paper. The remainder of the paper is organized as follows. In Section 2, we provide background on diffusion models and the RLHF framework. The theory is developed in

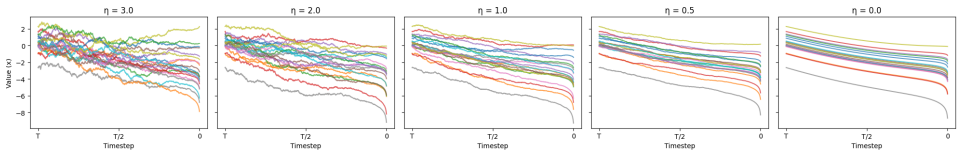

Figure 2: One-dimensional VP dynamics under different noise levels (η), using the same control function. Increasing η introduces higher stochasticity in the trajectories while preserving the underlying dynamics.

Sections 3 and 4. In Section 5, we conduct numerical experiments to justify the role of sampler stochasticity in fine-tuning large-scale T2I models. Concluding remarks are in Section 6.

2 Preliminaries

2.1 Diffusion Models

We review continuous-time diffusion models (Song et al., 2021b), which provide a unified treatment encompassing earlier discrete-time models. We follow the presentation in (Tang & Zhao, 2025).

The goal of diffusion models is to generate new samples that resemble the target data distribution $p_{\text{data}}(\cdot)$. It relies on a forward-backward procedure: the forward process is given by an SDE:

$$dX_t = f(t, X_t)dt + g(t)dW_t, \quad X_0 \sim p_{\text{data}}(\cdot), \tag{1}$$

where $\{W_t\}$ is d-dimensional Brownian motion, and $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ are model parameters. Some conditions on $f(\cdot, \cdot)$, $g(\cdot)$ are required so that equation 1 is well-defined, see (Stroock & Varadhan, 1979). We also assume that the forward process $\{X_t\}$ has a smooth density $p(t, x) := \mathbb{P}(X_t \in dx)/dx$. The success of continuous-time diffusion models is that their time reversal has a tractable form, making it more convenient for methodological development. It is known (Anderson, 1982; Haussmann & Pardoux, 1986) that the distribution of the time reversal process $\overline{X}_t = X_{T-t}$ is governed by the SDE: $d\overline{X}_t = (-f(T-t, \overline{X}_t) + g^2(T-t)\nabla \log p(T-t, \overline{X}_t))dt + g(T-t)dB_t$, $\overline{X}_0 \sim p(T, \cdot)$, where $\{B_t\}$ is a copy of d-dimensional Brownian motion.

There are two takeaways in diffusion modeling. First, diffusion models aim to generate the target distribution from *noise*, and the noise should not depend on the target distribution. So instead of setting $\overline{X}_0 \sim p(T, \cdot)$, the backward process is initiated with a prior $p_{\text{noise}}(\cdot)$ (as a proxy of $p(T, \cdot)$). The diffusion model is specified by the pair $(f(\cdot, \cdot), g(\cdot))$, and notable examples include variance exploding (VE) and variance preserving (VP) SDEs (see Section 3). The choice of $p_{\text{noise}}(\cdot)$ depends on each specific model, and varies case by case. Second, all but the term $\nabla \log p(T-t, \overline{X}_t)$ for the time-reversal process $\{\overline{X}_t\}$ are available. So we need to compute $\nabla \log p(t, x)$, known as *score function*. Score matching techniques allow to estimate the score function via a family of parameterized functions $\{s_\phi(t, x)\}_\phi$ (e.g., neural nets). The most widely used approach is *denoising score matching* Vincent (2011): $\min_\phi \mathbb{E}_{t\sim\text{Unif}\,[0,T],\, X_0\sim p_{data}(\cdot)}[\mathbb{E}_{p(t,\cdot|X_0)}|s_\phi(t, X_t) - \nabla \log p(t, X_t|X_0)|^2]$.

With the (true) score function $\nabla \log p(t, x)$ being replaced with the score matching function $s_\theta(t, x)$, the (backward) diffusion sampler is:

$$dY_t = \left(-f(T-t, Y_t) + g^2(T-t)s_\phi(T-t, Y_t)\right)dt + g(T-t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{2}$$

Equivalently, the sampler is set to:

$$dY_t = \left(-f(T-t, Y_t) + \frac{1+\eta^2}{2}g^2(T-t)s_\phi(T-t, Y_t)\right)dt + \eta\, g(T-t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot), \tag{3}$$

where $\eta \geq 0$ controls the *sampler stochasticity* along the trajectories. When $\eta = 1$, equation 3 specializes to equation 2; when $\eta = 0$, equation 3 becomes the probability flow ODE. The processes specified by equation 3 will be used as the *pretrained models* for fine-tuning.

2.2 Discretization – Stochastic gDDIM

Implementation of equation 3 requires suitable discretization. As shown in (Song et al., 2021b), the widely adopted DDPM is a discretized VP-SDE, and deterministic DDIM is discretization of the

probability flow VP-ODE. For most existing diffusion models, the model parameters have the form:

$$f(t,x) = \frac{1}{2}\frac{d\log\alpha_t}{dt}x \quad \text{and} \quad g(t) = \sqrt{-\frac{d\log\alpha_t}{dt}},$$

where α_t is a decreasing function with $\alpha_0 = 1$ and $\alpha_T = 0$. (Zhang et al., 2023) proposed the following discretization of equation 3 for all $\eta \geq 0$:

$$x_{t-\Delta t} = \sqrt{\frac{\alpha_{t-\Delta t}}{\alpha_t}}x_t + \left(\sqrt{\frac{\alpha_{t-\Delta t}}{\alpha_t}}(1-\alpha_t) - \sqrt{(1-\alpha_{t-\Delta t}-\sigma_t^2)(1-\alpha_t)}\right)s_\theta(t,x_t) + \sigma_t(\eta)\mathcal{N}(0,I), \quad x_T \sim p_{\text{noise}}(\cdot). \tag{4}$$

where

$$\sigma_t(\eta) = (1-\alpha_{t-\Delta t})\left(1 - \left(\frac{1-\alpha_{t-\Delta t}}{1-\alpha_t}\right)^{\eta^2}\left(\frac{\alpha_t}{\alpha_{t-\Delta t}}\right)^{\eta^2}\right). \tag{5}$$

The equation 4 is referred to as the *generalized denoising diffusion implicit model* (gDDIM), which agrees with stochastic DDIM (Song et al., 2021a). The difference lies in the choice of $\sigma_t(\eta)$: (Song et al., 2021a) relied on heuristics to suggest $\sigma_t(\eta) \propto \eta$, whereas equation 5 proposed by (Zhang et al., 2023) guarantees the exact sampling if the score function is known/precise. When $\eta = 0$, we have $\sigma_t(0) = 0$, and equation 4 is just deterministic DDIM.

2.3 DIFFUSION RLHF

RL provides a framework to fine-tune generative models in the context of RLHF (Bai et al., 2022; Ouyang et al., 2022), which was originally proposed for LLM alignment. Such a paradigm can also be made well-suited for fine-tuning diffusion models, particularly to enhance T2I generation aligned with human feedback. The first works in this direction are (Black et al., 2024; Fan et al., 2023; Lee et al., 2023), which formulated the denoising step as Markov decision processes (MDPs). (Gao et al., 2024; Zhao et al., 2024; 2025) developed an RL approach tailored to continuous-time models. Here we focus on the most widely used *Denoising Diffusion Policy Optimization* (DDPO).

DDPO (Black et al., 2024): We formulate the denoising steps $\{x_T, x_{T-1}, \cdots, x_0\}$ as an MDP:

$$s_t := (\boldsymbol{c}, t, x_t), \qquad \pi(a_t|s_t) := p_\theta(x_{t-1}|x_t,\boldsymbol{c}), \qquad \mathbb{P}(s_{t+1}|s_t,a_t) := \big(\delta_{\boldsymbol{c}}, \delta_{t-1}, \delta_{x_{t-1}}\big),$$

$$a_t := x_{t-1}, \qquad \rho_0(s_0) := \big(p(\boldsymbol{c}), \delta_T, \mathcal{N}(0,I)\big), \qquad R(s_t,a_t) := \begin{cases} r(x_0,\boldsymbol{c}) & \text{if } t=0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{c}$ is th prompt, and $\delta_\bullet$ is the Dirac mass on $\bullet$. Note that the reward function is only evaluated at the last denoising step on the generated image x_0. The objective of DDPO is to maximize:

$$\mathcal{J}_{DDPO}(\theta) := \mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}), x_0 \sim p_\theta(x_0|\boldsymbol{c})}[r(x_0,\boldsymbol{c})]. \tag{6}$$

The policy gradient of equation 6 is $\nabla_\theta \mathcal{J}_{DDPO} = \mathbb{E}\left[\sum_{t=0}^T \nabla_\theta \log p_\theta(x_{t-1}|x_t,\boldsymbol{c})\, r(x_0,\boldsymbol{c})\right]$ (Williams, 1992). It is common to parameterize $p_\theta(x_{t-1}|x_t,\boldsymbol{c})$ as isotopic Gaussian, so $\nabla_\theta \mathcal{J}_{DDPO}$ can be easily computed by Monte Carlo (Mohamed et al., 2020).

GRPO (Shao et al., 2024): A recent breakthrough in the space of RLHF is the *Group Relative Policy Optimization* (GRPO) framework in the wake of Deepseek-R1 for reasoning. For T2I tasks, GRPO uses the group normalization/average per prompt to compute the *advantages*:

$$A_i := \frac{r(x_0^i,\boldsymbol{c}) - \frac{1}{G}\sum_{k=1}^G r(x_0^k,\boldsymbol{c})}{\text{std}\left(\{r(x_0^k,\boldsymbol{c})\}_{k=1}^G\right)},$$

where G is the group size, and $\text{std}\left(\{r(x_0^k,\boldsymbol{c})\}_{k=1}^G\right)$ denotes the standard deviation of the group rewards. The objective of GRPO is to maximize:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\boldsymbol{c}\sim p(\boldsymbol{c}), \{x_t\}\sim\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{c})}\left[\frac{1}{G}\sum_{i=1}^G \frac{1}{T}\sum_{t=1}^T \min\left(\rho_t^i(\theta)A_i, \text{clip}(\rho_t^i(\theta), 1-\epsilon, 1+\epsilon)A_i\right)\right], \tag{7}$$

where $\rho_t^i(\theta) = \frac{p_\theta(x_{t-1}^i|x_t^i,\boldsymbol{c})}{p_{\theta_{\text{old}}}(x_{t-1}^i|x_t^i,\boldsymbol{c})}$ is the likelihood ratio. Several variants of the GRPO algorithm (Xue et al., 2025; Li et al., 2025) have also been proposed for T2I tasks.

3 THEORY ON THE REWARD GAP

As mentioned earlier, fine-tuning diffusion models via RL requires suitable stochasticity for exploration, while inference often relies on (efficient) deterministic samplers. This leads to a reward gap between the post-training and the inference processes. Here we will provide some theoretical insights on this reward gap in the continuous setting. Denote by $\{Y_t^{REF}\}$ the (reference) pretrained model specified by equation 3.

• *Fine-tuning*: Denote by $\{Y_t^{SDE}\} := \{Y_t^{\theta}\}$ the fine-tuned model parameterized by θ. We follow (Uehara et al., 2024; Tang, 2024) to use an entropy-regularized reward objective:

$$F_\eta(\theta) = \mathbb{E}[r(Y_T^\theta)] - \beta\,\mathrm{KL}(Y_T^\theta || Y_T^{REF}), \tag{8}$$

where $r(\cdot)$ is a reward function, and $\beta > 0$ controls the level of exploration relative to the pretrained model. The dependence on η comes from Y_T^{REF}, and the problem is to solve $\theta_\eta^* := \arg\max_\theta F_\eta(\theta)$.

• *Inference*: Denote by $\{Y_t^{ODE}\}$ the (deterministic) ODE sampler corresponding to $\{Y_t^{SDE}\}$ (by setting $\eta = 0$).

Also let $\mathcal{J}_\bullet := \mathbb{E}[r(Y_T^\bullet)]$ be the evaluation metric, with $\bullet \in \{REF, SDE, ODE\}$. The goal is to understand how much improvement gained from fine-tuning, and the reward gap induced by the (deterministic) ODE sampler. This motivates the following definitions:

Definition 3.1. *Define*

1. *Improvement of fine-tuning as the reward difference between $\{Y_t^{ODE}\}$ and $\{Y_t^{REF}\}$, $I_\eta := \mathcal{J}_{ODE} - \mathcal{J}_{REF}$.*
2. *Reward gap as the reward difference between $\{Y_t^{ODE}\}$ and $\{Y_t^{SDE}\}$, $\Delta_\eta := \mathcal{J}_{ODE} - \mathcal{J}_{SDE}$.*

In what follows, we study I_η and Δ_η for the VE/VP models with a Gaussian or mixture Gaussian target distribution, where the score function has a closed-form expression. A non-vacuous bound for the general case will be given in the next section.

3.1 VE WITH A GAUSSIAN TARGET

We first consider the one-dimensional VE model:

$$dX_t = \sqrt{2t}\,dB_t, \quad \text{with } p_{\text{data}}(\cdot) = \mathcal{N}(0,1). \tag{9}$$

Since $X_t \sim \mathcal{N}(0, t^2+1)$, the (exact) score function is $\nabla \log p(t,x) = -\frac{x}{t^2+1}$. So the "pretrained" model is:

$$dY_t^{REF} = -\frac{(1+\eta^2)(T-t)}{1+(T-t)^2} Y_t^{REF} dt + \eta\sqrt{2(T-t)} dB_t.$$

Next we set the reward function $r(x) = -(x-1)^2$, so the goal of fine-tuning is to drive the sample towards the mode 1. We also specify the fine-tuned SDE and ODE by

$$dY_t^{SDE} = -\frac{(1+\eta^2)(T-t)}{1+(T-t)^2}(Y_t^{SDE} + \theta_\eta^*)dt + \eta\sqrt{2(T-t)}dB_t,$$
$$dY_t^{ODE} = -\frac{T-t}{1+(T-t)^2}(Y_t^{ODE} + \theta_\eta^*)dt,$$

with θ_η^* maximizing the entropy-regularized reward (equation 8). The following theorem gives bounds for I_η and Δ_η under VE, and the proof is deferred to Appendix C.

Theorem 3.1. *Consider the VE model (equation 9), with the reward function*

$$r(x) = -(x-1)^2. \tag{10}$$

For $\eta > 0$, we have $\theta_\eta^* = -\left[(1+\frac{\beta}{2})\left(1-(1+T^2)^{-\frac{1+\eta^2}{2}}\right)\right]^{-1}$. Moreover,

$$\Delta_\eta \le \frac{1}{2T} + o\left(\frac{1}{T}\right) \quad \textit{and} \quad \mathcal{I}_\eta \ge 1 - \frac{1}{2T} + o\left(\frac{1}{T}\right). \tag{11}$$

3.2 VP WITH A GAUSSIAN TARGET

Now we consider the one-dimensional VP model:

$$dX_t = -tX_t dt + \sqrt{2t}\, dB_t, \quad \text{with } p_{\text{data}}(\cdot) = \mathcal{N}(0,1). \tag{12}$$

Under the same setup as in the VE case, we have:

$$\begin{aligned}
dY_t^{REF} &= -\eta^2(T-t)Y_t^{REF}dt + \eta\sqrt{2(T-t)}dB_t,\\
dY_t^{SDE} &= -\eta^2(T-t)Y_t^{SDE}\,dt - (1+\eta^2)(T-t)\theta_\eta^*(t)\,dt + \eta\sqrt{2(T-t)}dB_t,\\
dY_t^{ODE} &= -(T-t)\theta_\eta^*(t)dt,
\end{aligned}$$

where a time-dependent control $\theta_\eta(t) := \theta_\eta e^{-\frac{(T-t)^2}{2}}$ is used for fine-tuning. The following theorem gives bounds for I_η and Δ_η under VP, and the proof is deferred to Appendix xxx.

Theorem 3.2. *Consider the VP model (equation 12), with the reward function $r(x) = -(x-1)^2$. For $\eta > 0$, we have $\theta_\eta^* = -\left[(1+\frac{\beta}{2})\left(1 - e^{-\frac{(1+\eta^2)T^2}{2}}\right)\right]^{-1}$. Moreover,*

$$\Delta_\eta \le \frac{e^{-T^2}}{2} + o\left(e^{-T^2}\right) \quad \textit{and} \quad \mathcal{I}_\eta \ge 1 - \frac{e^{-T^2}}{2} + o\left(e^{-T^2}\right). \tag{13}$$

3.3 VE/VP WITH A MIXTURE GAUSSIAN TARGET

The previous results can be extended to multidimensional setting, with a mixture Gaussian target distribution. Recall that the probability density of a mixture Gaussian has the form:

$$\sum_{i=1}^{k} \frac{\alpha_i}{(2\pi)^{d/2}(\det \boldsymbol{\Sigma}_i)^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i (x-\boldsymbol{\mu}_i)\right),$$

where α_i is the weight of the i-th Gaussian component. The following corollary bounds the reward gap for a mixture Gaussian target distribution, which we will prove in Appendix C.

Corollary 3.1. *Let the reward function be $r(\boldsymbol{x}) = -||\boldsymbol{x} - \boldsymbol{r}||^2$ such that $\boldsymbol{\mu}_i \cdot \boldsymbol{r} = 0$ for all $i \in \{1,\dots,k\}$, $\Sigma_i \equiv \boldsymbol{I}_d$ and $\mathbb{E}[Y_T^{REF}] = \mathbf{0}$. Then the same bounds on reward gap hold, i.e.,*

$$\Delta_\eta \le \begin{cases} (2T)^{-1} & \textit{for VE},\\ e^{-T^2}/2 & \textit{for VP}. \end{cases} \tag{14}$$

For all these examples, the reward gap $\Delta_\eta \to 0$ (independent of η), as $T \to \infty$. Such a phenomenon will also be observed in fine-tuning T2I models with more complex rewards.

4 NON-VACUOUS BOUND ON THE REWARD GAP

In Section 3, we studied the reward gap for the VE and VP models with a Gaussian/mixture Gaussian target distribution. The analysis relies on the explicit computation of the score function, which is not available in general. Here our goal is to provide a bound on $W_2(Y_T^{SDE}, Y_T^{ODE})$, where

$$\begin{aligned}
dY_t^{SDE} &= \left(-f(T-t, Y_t^{SDE}) + \frac{1+\eta^2}{2} g^2(T-t) s_{\theta_\eta^*}(T-t, Y_t^{SDE})\right) dt + \eta\, g(T-t) dB_t,\\
dY_t^{ODE} &= \left(-f(T-t, Y_t^{ODE}) + \frac{1}{2} g^2(T-t) s_{\theta_\eta^*}(T-t, Y_t^{ODE})\right) dt.
\end{aligned}$$

Here we fine-tune directly on the score matching function $\{s_\theta(t,x)\}_\theta$, instead of adding a control as in Section 3. We need the following assumptions.

Assumption 4.1. *The following conditions hold:*

1. *Dissipativity of f: There is $m > 0$ such that $(y_2 - y_1)\cdot(-f(t,y_2) + f(t,y_1)) \le -m||y_2 - y_1||^2$ for all y_1, y_2.*

	η	ImageReward	PickScore	HPS_v2	Aesthetic
Base	–	0.320	20.69	0.253	5.375
ImageReward	1.0	0.837 (0.918)	20.89 (20.94)	**0.268** (**0.271**)	5.638 (5.720)
	1.2	**0.915** (**1.031**)	**20.95** (**20.99**)	**0.268** (0.289)	**5.697** (**5.803**)
	1.5	0.771 (0.907)	20.90 (20.95)	0.266 (0.269)	5.605 (5.735)
PickScore	1.0	0.587 (**0.729**)	21.11 (21.28)	**0.265** (**0.267**)	5.560 (5.678)
	1.2	**0.594** (0.692)	**21.17** (21.33)	0.264 (0.264)	5.568 (5.692)
	1.5	0.566 (0.701)	21.10 (**21.36**)	0.264 (0.262)	**5.599** (**5.769**)

Table 1: Performance for ODE (SDE) samplers under DDPO fine-tuning. Bold numbers indicate the highest evaluations among all stochasticity, demonstrating $\eta = 1.2$ in general performs the best.

2. *Lipschitz of s_θ: There is $L > 0$ such that $||s_\theta(t, y_1) - s_\theta(t, y_2)|| \leq L||y_1 - y_2||$ for all θ, y_1, y_2.*

3. *L^2 bound on Y^{SDE}: There is $A > 0$ such that $\sup_{0\leq t\leq T} \mathbb{E}[||Y_t^{SDE}||^2] \leq A$.*

The conditions 1 (dissipativity of f) and 2 (Lipschitz condition on s_θ) are standard in stability analysis. The condition 3 requires the process $\{Y_t^{SDE}\}$ be contractive, which holds if f is dissipative, and the fine-tuned distribution satisfies some strongly log-concave condition, see e.g., (Gao et al., 2025; Tang & Zhao, 2024). The following theorem gives a non-vacuous bound on the reward gap, which is proved in Appendix D.

Theorem 4.1. *Let Assumption 4.1 hold, and assume that $-2m + \left(L + \frac{1}{4}\eta^2\right)||g||_\infty^2 \leq -\kappa$ for some $\kappa > 0$. We have:*

$$W_2(Y_T^{SDE}, Y_T^{ODE}) \leq \eta L||g||_\infty \sqrt{\frac{A(1 - e^{-\kappa T})}{\kappa}}. \tag{15}$$

Moreover, if the reward function satisfies $|r(y_1) - r(y_2)| \leq C|y_1 - y_2|$ for some $C > 0$, then $\Delta_\eta \leq \eta CL||g||_\infty \sqrt{A\left(1 - e^{-\kappa T}\right)/\kappa}$.

The assumption on $-2m + \left(L + \frac{1}{4}\eta^2\right)||g||_\infty^2$, and the bound (equation 15) suggest that we avoid using an overly large stochasticity η to control the reward gap. This is also consistent with our empirical observation that the performance of T2I generation deteriorates when fine-tuning with very large η.

5 Numerical Experiments

In this section, we conduct numerical experiments for fine-tuning large-scale T2I models by leveraging sampler stochasticity, and examine the reward gap. Two classes of RL algorithms are considered: DDPO (Black et al., 2024) and MixGRPO (Li et al., 2025). The preference rewards that we use for fine-tuning DDPO include the LAION aesthetic (Schuhmann et al., 2022), HPS_v2.1 (Wu et al., 2023), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023). The rewards for fine-tuning MixGRPO include HPS_v2.1, Pick Score, ImageReward, and Unified Reward. (Wang et al., 2025) To study the reward gap, we use a high stochasticity η under gDDIM scheme to generate training samples, and compare them with the ODE samples generated using the same weights in each iteration.

5.1 DDPO

We use Stable Diffusion v1.5 (Rombach et al., 2022) as the base model, and fine-tune it with DDPO. During training, we adopt ImageReward and PickScore as preference rewards, while Aesthetic and HPS_v2 are included as additional evaluation metrics. Figure 1 (and Appendix F) provide representative generations from the fine-tuned models. Our main observations are summarized as follows:

• *Decreasing Reward Gap with Quality Improvement*: we conduct fine-tuning experiments under PickScore with $\eta = 1.2$ and calculate SDE–ODE reward (contray to Section 3, here we subtract ODE reward value from SDE reward for non-negativity) differences under multiple preference functions every 200 steps until reward collapse. As shown in Table 2, the gap decreases as image quality improves for both samplers.

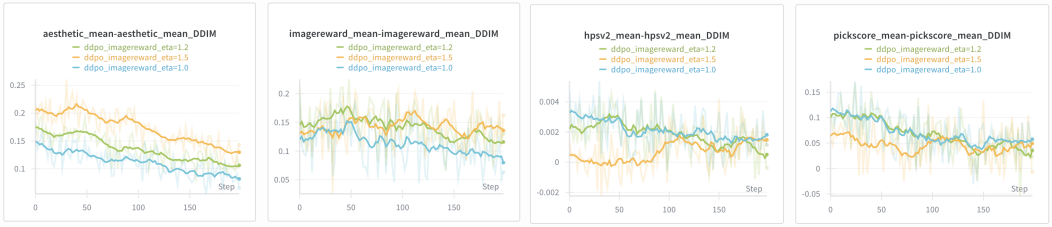

Figure 3: Evolution of reward gap during DDPO training under PickScore fine-tuning with stochasticity $\eta = 1.2$. The gap for multiple results decreases steadily as training progresses, indicating that image quality improves for both samplers and that ODE inference remains competitive.

• *High Stochasticity Benefits Moderate Time Steps*: we compare fine-tuning under ImageReward and PickScore at $\eta \in 1.0, 1.2, 1.5$ for 200 steps. As shown in Table 1, $\eta = 1.2$ under ImageReward achieves the best in-domain and out-of-domain performance, while PickScore's performances depend on evaluation metrics.

• *Richer Prompt Contents Reduce Reward Gap*: we compare performances with animal versus more comprehensive prompts (Figure 1, Appendix F) under ImageReward with $\eta = 1.2$. As shown in Table 3, complex prompts generate higher post-tuning rewards and higher in-group variance in post-training. Moreover, their richer instructions reduce the SDE–ODE reward gap.

	$T = 0$	**200**	**400**	**600**	**800**	**1000**	**1200**	**1400**	**(1476)**
ImgRwrd Gap	0.160	0.119	0.102	0.028	0.057	0.027	**0.006**	0.020	(0.564)
HPSv2 Gap	0.0047	0.0032	0.0032	0.0018	0.0027	0.0015	**-0.0028**	0.0011	(0.0308)
Aesthetic Gap	0.162	0.113	0.078	0.077	0.072	0.017	0.030	**-0.010**	(0.685)
PickScore Gap	0.115	0.146	0.167	0.118	0.178	0.106	0.129	**0.090**	(0.907)
SDE Reward	20.823	21.310	21.647	21.901	22.068	22.265	22.306	**22.417**	(18.920)
ODE Reward	20.730	21.165	21.500	21.783	21.883	22.172	22.195	**22.318**	(17.044)

Table 2: PickScore training until reward collapses. Smallest reward gaps and best sampler performances locate at the large training steps $T = 1200, 1400$.

	Animal Prompts			**Complicated Prompts**		
	$T = 0$	**100**	**200**	$T = 0$	**100**	**200**
Mean	**0.382** (**0.539**)	0.668 (0.805)	0.915 (1.031)	0.347 (0.470)	**0.763** (**0.872**)	**1.086** (**1.174**)
Std	0.814 (0.799)	0.775 (0.721)	0.709 (0.652)	**1.030** (**0.997**)	**0.909** (**0.868**)	**0.795** (**0.727**)
Gap	**0.146**	**0.138**	**0.116**	0.115	0.110	0.106

Table 3: Performance Comparison between prompts of different complexity. More complicated prompts yields faster fine-tuning improvements, larger in-group variances, and smaller SDE-ODE reward gaps.

5.2 MIXGRPO

We use FLUX.1 (Black Forest Labs, 2024) as the base model, and fine-tune it with MixGRPO, which is a sliding-window sampler that alternates between ODE and SDE schemes for 25 training steps in total. Training is carried out with multiple rewards combined using equal weights, while evaluation is reported on ImageReward and HPSClip.

• *Bounded Reward Gap Under High Stochasticity*: With $\eta = 1.2$, ODE sampling in MixGRPO consistently outperforms the mixed SDE–ODE scheme, in contrast to the results from DDPO. As shown in the top-left and bottom-left panels of Figure 4, the reward gap steadily diminishes during training and converges to zero.

• *Quality Improvement*: The middle and right panels of Figure 4 further shows that the ODE sampler performs consistently better on training prompts. For example, in the left pair of Figure 5, the

generation with $\eta = 1.2$ correctly aligns with the "trapped inside" prompt instruction, whereas the generation with $\eta = 1.0$ fails to do so.

Figure 4: Bounded reward gap (left column) and performance improvement (middle, right column) for MixGRPO

Figure 5: Comparison of ODE image generation by FLUX with MixGRPO fine-tuning, stochasticity $\eta = 1.2$ (below) and $\eta = 1.0$ (above). Higher stochasticity shows better alignments to details.
Prompts (from left to right): "A steampunk pocketwatch owl is trapped inside a glass jar buried in sand, surrounded by an hourglass and swirling mist.", "An androgynous glam rocker poses outside CBGB in the style of Phil Hale.", "A digital painting by Loish featuring a rush of half-body, cyberpunk androids and cyborgs adorned with intricate jewelry and colorful holographic dreads."

6 Conclusion and Further Directions

Sampling stochasticity in diffusion RLHF plays an important role yet remains underexamined. Here we present an explicit solution for generalized VE/VP diffusion models, together with extensive experiments, demonstrating that "high stochasticity in training samples, no stochasticity in generation" is both theoretically sound and practically advantageous. Our results open a broader space for tuning stochasticity hyperparameters, enabling more robust and diversified diffusion post-training. Future work will quantify the "reward gap" in video and multimodal generation and develop clearer theoretical insights into stochasticity as a drifting force toward the target distribution.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Process. Appl.*, 12(3): 313–326, 1982.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.

Black Forest Labs. Flux. *GitHub repository*, 2024. URL `https://github.com/black-forest-labs/flux`.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=zyLVMgsZ0U_`.

Jinyoung Choi, Junoh Kang, and Bohyung Han. Enhanced diffusion sampling via extrapolation with multiple ode solutions. *arXiv preprint arXiv:2504.01855*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *arXiv preprint arXiv:2409.08861*, 2024.

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2 edition, 2010.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Neurips*, volume 36, pp. 79858–79885, 2023.

Xuefeng Gao, Jiale Zha, and Xun Yu Zhou. Reward-directed score-based diffusion models via q-learning. *arXiv preprint arXiv:2409.04832*, 2024.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *J. Mach. Learn. Res.*, 26(43):1–54, 2025.

U. G. Haussmann and É. Pardoux. Time reversal of diffusions. *Ann. Probab.*, 14(4):1188–1205, 1986.

John R. Hershey and Peder A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pp. IV–317–IV–320, 2007. doi: 10.1109/ICASSP.2007.366913.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Fast convergence for high-order ode solvers in diffusion probabilistic models. *arXiv preprint arXiv:2506.13061*, 2025.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Kyungmin Lee, Xiaohang Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Lee_Calibrated_Multi-Preference_Optimization_for_Aligning_Diffusion_Models_CVPR_2025_paper.pdf.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. MixGRPO: Unlocking flow-based GRPO efficiency with mixed ODE-SDE. *arXiv preprint arXiv:2507.21802*, 2025.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, Mingda Wan, and Yufa Zhou. Unraveling the smoothness properties of diffusion models: A gaussian mixture perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 23–29 Jul 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

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

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pp. 1–22, 2025.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. Training language models to follow instructions with human feedback. In *Neurips*, volume 35, pp. 27730–27744, 2022.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Hannes Risken. *Fokker-Planck Equation*, pp. 63–95. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996. ISBN 978-3-642-61544-3. doi: 10.1007/978-3-642-61544-3_4.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Neurips*, volume 35, pp. 25278–25294, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv preprint arXiv:2209.13085*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Dmitrii Sorokin, Maksim Nakhodnov, Andrey Kuznetsov, and Aibek Alanov. Imagerefl: Balancing quality and diversity in human-aligned diffusion models. *arXiv preprint arXiv:2505.22569*, 2025.

Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*, volume 233 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 1979.

Wenpin Tang. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond. *arXiv preprint arXiv:2403.06279*, 2024.

Wenpin Tang and Hanyang Zhao. Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*, 2024.

Wenpin Tang and Hanyang Zhao. Score-based diffusion models via stochastic differential equations. *Stat. Surv.*, 19:28–64, 2025.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3-4):229–256, 1992.

Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D Yao, Shi-Xiong Zhang, and Sambit Sahu. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *J. Artif. Intell. Res.*, 82:2595–2661, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for Gaussian mixture models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53291–53327. PMLR, 21–27 Jul 2024.

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. In *Neurips*, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.

Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods. *arXiv preprint arXiv:2502.01384*, 2025.

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

Qinsheng Zhang, Molei Tao, and Yongxin Chen. gDDIM: generalized denoising diffusion implicit models. In *ICLR*, 2023.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David Yao, and Wenpin Tang. Scores as Actions: a framework of fine-tuning diffusion models by continuous-time reinforcement learning. *arXiv preprint arXiv:2409.08400*, 2024.

Hanyang Zhao, Haoxian Chen, Ji Zhang, David Yao, and Wenpin Tang. Score as Action: Fine tuning diffusion generative models by continuous-time reinforcement learning. In *ICML*, 2025.

A LEMMAS ON LINEAR DYNAMICS WITH GAUSSIAN PRIORS

Lemma A.1. *A stochastic process $\{Z_t\}_{t=0}^T$ with first order linear dynamic and initial Gaussian distribution*

$$\begin{cases} dZ_t = f(t)Z_t dt + g(t)dt + h(t)dB_t \quad t \in [0,T] \\ Z_0 \sim \mathcal{N}(0,1) \end{cases}$$

is distributed following

$$Z_t \sim \mathcal{N}\Big(e^{F(t)} \int_0^t e^{-F(s)} g(s)\, ds,\ e^{2F(t)} \int_0^t e^{-2F(t)} h^2(s)\, ds\Big),$$

in which $F(t)$ is the integrating factor satisfying $F(t) = \int_0^t f(s)ds$.

Lemma A.2. *A stochastic process $\{Z_t\}_{t=0}^T$ with first order linear dynamic and initial Gaussian distribution*

$$\begin{cases} dZ_t = f(t)Z_t dt + g(t)dB_t \quad t \in [0,T] \\ Z_0 \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \end{cases}$$

is distributed following

$$\begin{aligned} Z_t &\sim Z_0 \cdot e^{F(t)} + \mathcal{N}\Big(0, \int_0^t e^{-2F(s)} g^2(s) ds\Big) \cdot e^{F(t)} \\ &\sim \mathcal{N}\Big(\mu_Z \cdot e^{F(t)}, \Big[\sigma_Z^2 + \int_0^t e^{-2F(s)} g^2(s) ds\Big] \cdot e^{2F(t)}\Big), \end{aligned}$$

in which $F(t)$ is the integrating factor satisfying $F(t) = \int_0^t f(s)ds$.

Lemma A.3. *A parametrized family of stochastic processes $\{Z_t^\theta\}_{t=0}^T$ with initial Gaussian distribution*

$$\begin{cases} dZ_t^\theta = f(t) \cdot (Z_t^\theta + \theta(t))dt + g(t)dB_t \quad t \in [0,T] \\ Z_0 \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \end{cases}$$

is distributed following

$$\begin{aligned} Z_t &\sim Z_0 \cdot e^{F(t)} + \mathcal{N}\Big(\int_0^t e^{-F(s)} f(s)\theta(s) ds, \int_0^t e^{-2F(s)} g^2(s) ds\Big) \cdot e^{F(t)} \\ &\sim \mathcal{N}\Big(\mu_Z \cdot e^{F(t)} + \int_0^t e^{-F(s)} f(s)\theta(s) ds, \Big[\sigma_Z^2 + \int_0^t e^{-2F(s)} g^2(s) ds\Big] \cdot e^{2F(t)}\Big), \end{aligned}$$

in which $F(t)$ is the integrating factor satisfying $F(t) = \int_0^t f(s)ds$.

B USEFUL PROPOSITIONS

Proposition B.1. *The terminal distribution of Variance Exploding parametrized backward dynamics* $\{Y_t^\theta\}_{t=0}^T$ *is always Gaussian following the law*

$$\mathcal{N}(\mu_{Y_T^\theta}, \sigma^2_{Y_T}) := \mathcal{N}\Big(\theta \cdot (1+T^2)^{-\frac{1+\eta^2}{2}} - \theta, 1-(1+T^2)^{-(1+\eta^2)}\Big).$$

Therefore, reward function (10) takes the form of

$$\mathcal{J}_\eta(\theta) = -\Big(\sigma^2_{Y_T} + (1-\mu_{Y_T^\theta})^2\Big).$$

Proof. By comparing the coefficients with Lemma A.1, we have

$$\begin{cases} f_\eta(t) = -\frac{(1+\eta^2)(T-t)}{1+(T-t)^2} \\ g_\eta(t) = \eta\sqrt{2(T-t)} \\ \sigma_Z^2 = T^2. \end{cases}$$

Therefore, we first examine the exponential of integrating factor,

$$\begin{aligned} e^{F_\eta(t)} &= \exp\left(\int_0^t f_\eta(s)ds\right) \\ &= \exp\left(\int_0^t -\frac{(1+\eta^2)(T-s)}{1+(T-s)^2}ds\right) \\ &= \exp\left(\int_{1+T^2}^{1+(T-t)^2} \frac{(1+\eta^2)d(1+(T-s)^2)}{2(1+(T-s)^2)}\right) \\ &= \Big(\frac{1+T^2}{1+(T-t)^2}\Big)^{-\frac{1+\eta^2}{2}}. \end{aligned}$$

At terminal time T, the cumulative factor is

$$e^{F_\eta(T)} = \exp\left(\int_0^T f_\eta(s)ds\right) = \big(1+T^2\big)^{-\frac{1+\eta^2}{2}}.$$

Also, the cumulative Gaussian variance generated from the backward process is,

$$\begin{aligned} \int_0^t e^{-2F_\eta(s)} g_\eta^2(s)ds &= \int_0^t \Big(\frac{1+T^2}{1+(T-s)^2}\Big)^{(1+\eta^2)} \cdot (2\eta^2(T-s))ds \\ &= \int_{1+T^2}^{1+(T-t)^2} (-\eta^2)\Big(\frac{1+T^2}{1+(T-s)^2}\Big)^{(1+\eta^2)} d(1+(T-s)^2) \\ &= (1+T^2)^{(1+\eta^2)} \int_{1+T^2}^{1+(T-t)^2} (-\eta^2)(1+(T-s)^2)^{-(1+\eta^2)}\, d(1+(T-s)^2) \\ &= (1+T^2)^{(1+\eta^2)} \cdot \Big((1+(T-t)^2)^{-\eta^2} - (1+T^2)^{-\eta^2}\Big). \end{aligned}$$

At terminal time T, the variance from the process is

$$\begin{aligned} \int_0^T e^{-2F_\eta(s)} g_\eta^2(s)ds &= (1+T^2)^{(1+\eta^2)} \cdot \Big(1-(1+T^2)^{-\eta^2}\Big) \\ &= (1+T^2)^{(1+\eta^2)} - (1+T^2). \end{aligned}$$

Together with the initial Gaussian variance, the terminal distribution remains a zero-mean Gaussian:

$$\begin{aligned} Y_T &\sim \mathcal{N}\Big(0, \big[(1+T^2)^{(1+\eta^2)} - (1+T^2-\sigma_Z)\big] \cdot (1+T^2)^{-(1+\eta^2)}\Big) \\ &\sim \mathcal{N}\Big(0, \big[(1+T^2)^{(1+\eta^2)} - 1\big] \cdot (1+T^2)^{-(1+\eta^2)}\Big) \\ &\sim \mathcal{N}\Big(0, 1-(1+T^2)^{-(1+\eta^2)}\Big). \end{aligned}$$

Now we consider the terminal distribution for the parametrized process Y_T^θ. To reduce the problem to a dynamic with linear drift term, we define

$$Z_t^\theta = Y_t^\theta + \theta.$$

so that

$$\begin{cases} dZ_t = f_\eta(t)Z_t dt + g_\eta(t)dB_t \quad t \in [0,T] \\ Z_0 \sim \mathcal{N}(\theta, T^2) \end{cases}$$

Observe that both the dynamic variance and the initial distribution variance for $\{Z_t^\theta\}$ and $\{Y_t\}$ are the same, we may directly apply Lemma A.1 to obtain

$$Z_T^\theta \sim \mathcal{N}\big(\theta \cdot (1+T^2)^{-\frac{1+\eta^2}{2}}, 1-(1+T^2)^{-(1+\eta^2)}\big)$$

Therefore,

$$Y_T^\theta \sim \mathcal{N}\big(\theta \cdot (1+T^2)^{-\frac{1+\eta^2}{2}} - \theta, 1-(1+T^2)^{-(1+\eta^2)}\big),$$

and thus

$$\begin{cases} \mu_{Y_T^\theta} := \theta \cdot (1+T^2)^{-\frac{1+\eta^2}{2}} - \theta \\ \sigma^2_{Y_T} := 1-(1+T^2)^{-(1+\eta^2)} \end{cases} \tag{16}$$

In addition, we can now give a closed form representation of our reward,

$$\begin{aligned} \mathcal{J}_\eta(\theta) &= \mathbb{E}[(Y_T^\theta - 1)^2] \\ &= \mathbb{E}[Y_T^2] - 2\mathbb{E}[Y_T] + 1 \\ &= \sigma^2_{Y_T} + \mu^2_{Y_T^\theta} - 2\mu_{Y_T^\theta} + 1 \\ &= \sigma^2_{Y_T} + \big(1 - \mu_{Y_T^\theta}\big)^2, \end{aligned}$$

which yields to the desired expression. □

Proposition B.2. *Given β, η, T, the unique maximizer to the entropy regularized target (8) is:*

$$\theta_\eta^* = -\Big((1+\frac{\beta}{2}) \cdot \big[1-(1+T^2)^{-\frac{1+\eta^2}{2}}\big]\Big)^{-1}.$$

Proof. A classical distance result on two Gaussian distributions (Hershey & Olsen, 2007) states:

Lemma B.1. *The KL divergence for two Gaussian distributions $P \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Q \sim \mathcal{N}(\mu_2, \sigma_2)$,*

$$\mathrm{KL}(P||Q) = \log\Big(\frac{\sigma_2}{\sigma_1}\Big) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

In our model, $P \sim Y_T^\theta$ and $Q \sim Y_T^\odot$, so $\mu_1 = \mu_{Y_T^\theta}$, $\sigma_1 = \sigma_{Y_T}$, $\mu_2 = 0$, $\sigma_2 = 1$.

$$\begin{aligned} \mathrm{KL}(Y_T^\theta || Y_T^\odot) &= \log(\frac{1}{\sigma_{Y_T}}) + \frac{\sigma^2_{Y_T} + \mu^2_{Y_T^\theta} - 1}{2} \\ &= -\log(\sigma_{Y_T}) + \frac{\sigma^2_{Y_T} + \mu^2_{Y_T^\theta} - 1}{2}. \end{aligned}$$

Therefore,

$$\begin{aligned} -F_\eta(\theta) &= \sigma^2_{Y_T} + \big(\mu_{Y_T^\theta} - 1\big)^2 + \big(-\beta\log(\sigma_{Y_T}) + \frac{\beta}{2} \cdot (\sigma^2_{Y_T} + \mu^2_{Y_T^\theta} - 1)\big) \\ &= \big(\sigma^2_{Y_T} - \beta\log(\sigma_{Y_T}) + \frac{\beta}{2}\sigma^2_{Y_T} - \frac{\beta}{2}\big) + \big(\mu_{Y_T^\theta} - 1\big)^2 + \frac{\beta}{2}\mu^2_{Y_T^\theta} \\ &= \big(\sigma^2_{Y_T} - \beta\log(\sigma_{Y_T}) + \frac{\beta}{2}\sigma^2_{Y_T} - \frac{\beta}{2} + 1\big) + \big(1+\frac{\beta}{2}\big)\mu^2_{Y_T^\theta} - 2\mu_{Y_T^\theta}. \end{aligned}$$

So it suffices to minimize a quadratic function w.r.t $\mu_{Y_T^\theta}$, of which we know

$$\mu_{Y_T^{\theta^*}} = -\frac{-2}{2(1+\frac{\beta}{2})} = (1+\frac{\beta}{2})^{-1};$$

and thus

$$\theta_\eta^* \cdot \left[(1+T^2)^{-\frac{1+\eta^2}{2}} - 1\right] = \mu_{Y_T^{\theta^*}} = (1+\frac{\beta}{2})^{-1}.$$

This gives us the unique maximizer

$$\theta_\eta^* = \left((1+\frac{\beta}{2}) \cdot \left[(1+T^2)^{-\frac{1+\eta^2}{2}} - 1\right]\right)^{-1}$$

as desired. □

Proposition B.3. *The terminal distribution of Variance Preserving parametrized backward dynamics $\{Y_t^\theta\}_{t=0}^T$ is always Gaussian following the law*

$$\mathcal{N}(\mu_{Y_T^\theta}, 1) := \mathcal{N}\Big(\theta \cdot e^{-\frac{(1+\eta^2)\cdot T^2}{2}} - \theta, 1\Big).$$

And reward function (10) takes the form of $\mathcal{J}_\eta(\theta) = -\Big(1 + (1-\mu_{Y_T^\theta})^2\Big)$.

Proof. By comparing with the coefficients in Lemma A.2, we have

$$\begin{cases} f_\eta(t) = -\eta^2(T-t) \\ g_\eta(t) = -(1+\eta^2)(T-t)e^{-\frac{(T-t)^2}{2}}\theta_\eta^* \\ h_\eta(t) = \eta\sqrt{2(T-t)} \end{cases}$$

Therefore, we first examine the exponential of integrating factor,

$$\begin{aligned} e^{F_\eta(t)} &= \exp\left(\int_0^t f_\eta(s)ds\right) \\ &= \exp\left(-\eta^2 \int_{T-t}^T s\, ds\right) \\ &= \exp\left(-\frac{\eta^2}{2} \cdot \big(T^2 - (T-t)^2\big)\right). \end{aligned}$$

At terminal time T, the cumulative factor is

$$e^{F_\eta(T)} = \exp\left(-\frac{\eta^2}{2} \cdot \big(T^2 - (T-T)^2\big)\right) = e^{-\frac{\eta^2 T^2}{2}}.$$

Now we are able to compute

$$\begin{aligned} \int_0^t e^{-F_\eta(s)} g(s)\, ds &= \int_0^t e^{\frac{\eta^2}{2}(T^2-(T-s)^2)} \cdot \left(-(1+\eta^2)(T-s)e^{-\frac{(T-s)^2}{2}}\theta_\eta^*\right) ds \\ &= (1+\eta^2)\cdot\theta_\eta^* \cdot \int_{T-t}^T e^{\frac{\eta^2}{2}(T^2-s^2)} \cdot \left(se^{-\frac{s^2}{2}}\right) ds \\ &= (1+\eta^2)\cdot\theta_\eta^* \cdot \int_{T-t}^T e^{\frac{\eta^2 T^2}{2}} \cdot \left(se^{-\frac{(1+\eta^2)s^2}{2}}\right) ds \\ &= (1+\eta^2)\cdot\theta_\eta^* \cdot e^{\frac{\eta^2 T^2}{2}} \cdot \left[\frac{1}{1+\eta^2} \cdot e^{-\frac{(1+\eta^2)s^2}{2}}\right]_{s=T-t}^{s=T} \\ &= \theta_\eta^* \cdot e^{\frac{\eta^2 T^2}{2}} \cdot \left(e^{-\frac{(1+\eta^2)T^2}{2}} - e^{-\frac{(1+\eta^2)(T-t)^2}{2}}\right) \end{aligned}$$

Therefore,

$$\mu_{Y_T^\theta} = e^{-\frac{\eta^2 T^2}{2}} \cdot \theta_\eta^* \cdot e^{\frac{\eta^2 T^2}{2}} \cdot (e^{-\frac{(1+\eta^2)T^2}{2}} - e^0) = \theta_\eta^* \cdot (e^{-\frac{(1+\eta^2)T^2}{2}} - 1)$$

To the variance preserving property, it suffices to show

$$\frac{1}{e^{2F_\eta(t)}} = \int_0^t e^{-2F_\eta(t)} h_\eta^2(s)\, ds.$$

In fact,

$$\frac{d}{dt} e^{-2F_\eta(t)} = e^{-2F_\eta(t)} \cdot \frac{d(-\eta^2(T-t)^2)}{dt} = e^{-2F_\eta(t)} \cdot \big(\eta^2 \cdot 2(T-t)\big) = e^{-2F_\eta(t)} \cdot h_\eta^2(t).$$

□

Proposition B.4. *Given β, η, T, the unique maximizer to the entropy regularized target (8) is:*

$$\theta_\eta^* = -\Big((1+\frac{\beta}{2}) \cdot \big[1 - e^{-\frac{(1+\eta^2)\cdot T^2}{2}}\big]\Big)^{-1}.$$

Proof. Since $\sigma \equiv 1$, according to Lemma B.1, the maximum reward is attained at

$$\mu_{Y_T^{\theta^*}} = -\frac{-2}{2(1+\frac{\beta}{2})} = (1+\frac{\beta}{2})^{-1}.$$

By Proposition B.1,

$$\theta_\eta^* \cdot \big[e^{-\frac{(1+\eta^2)\cdot T^2}{2}} - 1\big] = \mu_{Y_T^{\theta^*}} = (1+\frac{\beta}{2})^{-1}.$$

This gives us the unique maximizer

$$\theta_\eta^* = \Big((1+\frac{\beta}{2}) \cdot \big[e^{-\frac{(1+\eta^2)\cdot T^2}{2}} - 1\big]\Big)^{-1}.$$

□

C Proofs for Section 3

C.1 Proof for Theorem 3.1

We first consider the Y_T^{ODE} process as discussed in Section 3.1. Since $\eta = 0$, by Proposition B.2,

$$\mu_{Y_T^{ODE}} = \theta_\eta^* \cdot (1+T^2)^{-\frac{1}{2}} - \theta_\eta^* = (1+\frac{\beta}{2})^{-1} \cdot \frac{(1+T^2)^{-\frac{1}{2}} - 1}{(1+T^2)^{-\frac{1+\eta^2}{2}} - 1},$$

and

$$\sigma^2_{Y_T^{ODE}} = \sigma^2_{Y_T^0} = 1 - (1+T^2)^{-1}.$$

Moreover, the quadratic reward $\mathcal{J}_{ODE}$ for Y_T^{ODE} is

$$\begin{aligned}\mathcal{J}_0(\theta_\eta^*) &= -\big(\sigma^2_{Y_T^{ODE}} + (\mu_{Y_T^{ODE}} - 1)^2\big) \\ &= (-1) + (1+T^2)^{-1} - \left(1 - (1+\frac{\beta}{2})^{-1} \cdot \frac{(1+T^2)^{-\frac{1}{2}} - 1}{(1+T^2)^{-\frac{1+\eta^2}{2}} - 1}\right)^2.\end{aligned}$$

Similarly, reward $\mathcal{J}_{SDE}$ for Y_T^{SDE} is

$$\begin{aligned}\mathcal{J}_\eta(\theta_\eta^*) &= -\big(\sigma^2_{Y_T^{SDE}} + (\mu_{Y_T^{SDE}} - 1)^2\big) \\ &= (-1) + (1+T^2)^{-(1+\eta^2)} - \Big(1 - (1+\frac{\beta}{2})^{-1}\Big)^2.\end{aligned}$$

For simplification, we denote

$$\bar{T} := 1 + T^2 \in (T^2, 2T^2), \quad \bar{\beta} := (1+\frac{\beta}{2})^{-1} \in (0, 1].$$

Now the reward gap

$$\begin{aligned}\Delta_\eta &= \mathcal{J}_{SDE} - \mathcal{J}_{ODE} \\ &= \Big((1+T^2)^{-(1+\eta^2)} - (1+T^2)^{-1}\Big) \\ &\quad - \left(\Big(1 - (1+\frac{\beta}{2})^{-1}\Big)^2 - \left(1 - (1+\frac{\beta}{2})^{-1} \cdot \frac{(1+T^2)^{-\frac{1}{2}} - 1}{(1+T^2)^{-\frac{1+\eta^2}{2}} - 1}\right)^2\right) \\ &= \left(\frac{\bar{T}^{-\eta^2} - 1}{\bar{T}}\right) - \left((1-\bar{\beta})^2 - \left(1 - \bar{\beta} \cdot \frac{\bar{T}^{-\frac{1}{2}} - 1}{\bar{T}^{-\frac{1+\eta^2}{2}} - 1}\right)^2\right).\end{aligned}$$

Since $0 < \bar{T}^{-\eta^2} \leq 1$, we can bound

$$-\bar{T}^{-1} < \frac{\bar{T}^{-\eta^2} - 1}{\bar{T}} \leq 0\,,$$

and

$$1 - \bar{T}^{-\frac{1}{2}} = \frac{\bar{T}^{-\frac{1}{2}} - 1}{0 - 1} < \frac{\bar{T}^{-\frac{1}{2}} - 1}{\bar{T}^{-\frac{1+\eta^2}{2}} - 1} \leq \frac{\bar{T}^{-\frac{1}{2}} - 1}{\bar{T}^{-\frac{1}{2}} - 1} = 1.$$

Therefore,

$$\begin{aligned}|\Delta_\eta| &= \left|\frac{\bar{T}^{-\eta^2} - 1}{\bar{T}}\right| + \left|(1-\bar{\beta})^2 - \left(1 - \bar{\beta} \cdot \frac{\bar{T}^{-\frac{1}{2}} - 1}{\bar{T}^{-\frac{1+\eta^2}{2}} - 1}\right)^2\right| \\ &\leq \bar{T}^{-1} + \left|(1-\bar{\beta})^2 - \Big((1-\bar{\beta}) + \bar{\beta} \cdot \bar{T}^{-\frac{1}{2}}\Big)^2\right| \\ &\leq \bar{T}^{-1} + \Big(\bar{\beta}^2 \cdot \bar{T}^{-1} + 2\bar{\beta}(1-\bar{\beta})\bar{T}^{-\frac{1}{2}}\Big) \\ &\leq 2\bar{\beta}(1-\bar{\beta})T^{-1} + o(T^{-1}) \\ &\leq \frac{1}{2T} + o\left(\frac{1}{T}\right).\end{aligned}$$

Now we attempt to bound the reference gap. The reward $\mathcal{J}_{REF}$ for Y_T^{REF} is

$$\begin{aligned}\mathcal{J}_\eta(0) &= -\big(\sigma^2_{Y_T^{REF}} + (\mu_{Y_T^{REF}} - 1)^2\big) \\ &= (-1) + (1+T^2)^{-(1+\eta^2)} - 1.\end{aligned}$$

Finally we have,

$$\begin{aligned}\Delta_\eta^{REF} &= \mathcal{J}_{ODE} - \mathcal{J}_{REF} \\ &= \bar{T}^{-1} - \left(1 - \bar{\beta}\cdot\frac{\bar{T}^{-\frac{1}{2}} - 1}{\bar{T}^{-\frac{1+\eta^2}{2}} - 1}\right)^2 - \left(\bar{T}^{-(1+\eta^2)} - 1\right) \\ &\geq \bar{T}^{-1} - \left(2\bar{\beta}(1-\bar{\beta})\bar{T}^{-\frac{1}{2}} + o(\bar{T}^{-\frac{1}{2}})\right) - \left(\bar{T}^{-1} - 1\right) \\ &\geq 1 - \frac{1}{2T} + o\left(\frac{1}{T}\right)\end{aligned}$$

C.2 Proof for Theorem 3.2

Similar to Appendix C.1, the quadratic reward (10) for Y_T^{ODE} is

$$\begin{aligned}\mathcal{J}_{ODE} &= -\big(\sigma^2_{Y_T^{ODE}} + (\mu_{Y_T^{ODE}} - 1)^2\big) \\ &= (-1) - \left(1 - (1+\frac{\beta}{2})^{-1}\cdot\frac{e^{-\frac{T^2}{2}} - 1}{e^{-\frac{(1+\eta^2)T^2}{2}} - 1}\right)^2.\end{aligned}$$

And the reward for Y_T^{SDE} is

$$\begin{aligned}\mathcal{J}_{SDE} &= -\big(\sigma^2_{Y_T^{SDE}} + (\mu_{Y_T^{SDE}} - 1)^2\big) \\ &= (-1) - \left(1 - (1+\frac{\beta}{2})^{-1}\right)^2.\end{aligned}$$

With $\bar{\beta} := (1+\frac{\beta}{2})^{-1} \in (0,1]$,

$$\begin{aligned}|\Delta_\eta| &= \left|(1-\bar{\beta})^2 - \left(1 - \bar{\beta}\cdot\frac{e^{-\frac{T^2}{2}} - 1}{e^{-\frac{(1+\eta^2)T^2}{2}} - 1}\right)^2\right| \\ &\leq \left|(1-\bar{\beta})^2 - \left((1-\bar{\beta}) + \bar{\beta}\cdot e^{-\frac{T^2}{2}}\right)^2\right| \\ &\leq \frac{e^{-\frac{T^2}{2}}}{2} + o\left(e^{-T^2}\right).\end{aligned}$$

Finally, $Y_T^{REF} \sim \mathcal{N}(0,1)$, so

$$\mathcal{J}_{REF} = (-1) - (1-0)^2.$$

And thus

$$\begin{aligned}\Delta_\eta^{REF} &= \mathcal{J}_{ODE} - \mathcal{J}_{REF} \\ &= 1 - \left(1 - (1+\frac{\beta}{2})^{-1}\cdot\frac{e^{-\frac{T^2}{2}} - 1}{e^{-\frac{(1+\eta^2)T^2}{2}} - 1}\right)^2 \\ &\geq 1 - \frac{e^{-\frac{T^2}{2}}}{2} + o(e^{-T^2})\end{aligned}$$

C.3 Proof for Proposition 3.1

We decompose $Y^{\theta,\bullet} := Y_{\parallel}^{\theta} + Y_{\perp}^{\theta}$ in terms of $\boldsymbol{r}$. Therefore,

$$r(Y^{\theta,\bullet}(T)) = -\left|Y_{\perp}^{\theta,\bullet} + (Y_{\parallel}^{\theta,\bullet} - \boldsymbol{r})\right|^2 = -\left(\left|Y_{\perp}^{\theta,\bullet}\right|^2 + \left|(Y_{\parallel}^{\theta,\bullet} - \boldsymbol{r})\right|^2\right)$$

Since $\boldsymbol{\mu}_i \perp \boldsymbol{r}$, $\mathbb{E}[Y_{\parallel}^{\theta}(0)] = 0$. Also, the reverse dynamics inject a scaled multiple of $\boldsymbol{I}_d$ noise, so we may analyze the Span($\boldsymbol{r}$) subspace separately, in which $\{Y_{\parallel}^{\theta_{\parallel}^*,SDE}\}$ and $\{Y_{\parallel}^{\theta_{\parallel}^*,ODE}\}$ follows a similar dynamics as discussed in Theorem 3.1 and Theorem 3.2, since the score function can be bounded according to (Liang et al., 2025).

On the other hand, $\theta_{\perp} = \mathbf{0}$ is a minimizer for $\left|Y_{\perp}^{\theta,\bullet}\right|^2$, since a non-trivial drift perpendicular to $\boldsymbol{r}$ does not alter the varianace (VP) or to the order of $O(1/T)$ (VE) but pushes the final distribution away from $r_{\perp} = \mathbf{0}$. Therefore, $\theta = \theta_{\parallel}$ and a same analysis as in Theorem 3.1 and Theorem 3.2 may apply to the similar dynamics.

D Proof for Section 4

Proposition D.1. *(Gronwall's) If* $u'(t) \le \alpha(t)u(t) + \beta(t)$,

$$u(T) \le e^{-\alpha(T)}u(0) + \int_0^T e^{-\alpha(T-s)}b(s)ds$$

Proof. The usual Gronwall's theorem (Evans, 2010) gives

$$u(T) \le e^{\int_t^T \alpha(s)ds}\left(u(0) + \int_0^T e^{-\alpha(s)}\beta(s)ds\right).$$

When $\alpha(t) < 0$, the proposition follows naturally. □

Let $u(t) := \mathbb{E}\left[||Y_T^{ODE} - Y_T^{SDE}||^2\right]$. Also, $Y_0^{\bullet} \sim p_{noise}$ implies $u(0) = 0$. Therefore, with appropriate α, β that satisfies the Gronwall's conditions, we can bound

$$u(T) \le e^{\int_t^T \alpha(s)ds}\left(\int_0^T e^{-\alpha(s)}\beta(s)ds\right). \tag{17}$$

By Ito's Lemma and Young's Inequality,

$$\begin{aligned}
u'(t) &= \frac{d}{dt}\mathbb{E}\left[||Y_T^{ODE} - Y_T^{SDE}||^2\right] \\
&= 2\mathbb{E}\big\langle Y_t^{ODE} - Y_t^{SDE}, -f(t, Y_t^{SDE}) + f(t, Y_t^{ODE})\big\rangle \\
&\quad + g^2(t)\mathbb{E}\big\langle Y_t^{ODE} - Y_t^{SDE}, s(t, Y_t^{ODE}) - s(t, Y_t^{SDE}))\big\rangle \\
&\quad - \eta^2 g^2(t)\mathbb{E}\big\langle Y_t^{ODE} - Y_t^{SDE}, s(t, Y_t^{SDE}) + \theta(t)\big\rangle + \eta^2 g^2(t) \\
&\le -2m \cdot u(t) - g^2(t)(L \cdot u(t)) \\
&\quad - \eta^2 g^2(t)\left(\frac{1}{2}u(t) + \frac{1}{2}\mathbb{E}\left[||s(t, Y_t^{SDE}) + \theta(t)||^2\right]\right) + \eta^2 g^2(t) \\
&= (-2m - Lg^2(t) + \frac{\eta^2 g^2}{2}) \cdot u(t) + \eta^2 g^2(t)\left(\frac{1}{2}\mathbb{E}\left[||s(t, Y_t^{SDE}) + \theta(t)||^2\right] + 1\right) \\
&\le \underbrace{(-2m - Lg^2(t) + \frac{\eta^2 g^2}{2})}_{\alpha(t)} \cdot u(t) + \underbrace{\eta^2 g^2(t)\left(\frac{L^2A^2}{2} + 1\right)}_{\beta(t)},
\end{aligned}$$

in which the last inequality makes use assumption (2) and (3), so that

$$\mathbb{E}[||s(t, Y_t^{SDE}||^2] \leq L^2 \mathbb{E}[||Y_t^{SDE}||^2] \leq L^2 A^2.$$

Therefore, with equation 17,

$$\begin{aligned} u(T) &\lesssim \int_0^T e^{-\kappa(T-s)} b(s) ds \\ &\lesssim \frac{\eta^2 L^2 A^2}{\kappa} (1 - e^{-\kappa T}) \end{aligned}$$

E LLM USAGE

Large Language Model (LLM) assists in LaTeX graphic alignments, spelling checks, and solving environment conflict issues in implementing DDPO and MixGRPO.

F More Experiment Results

F.1 DDPO Images on different training steps

Figure 6: SDE (above) and ODE (below) samples every 100 training steps under PickScore.

F.2 DDPO Images under ImageReward

Figure 7: Rows (from left to right): (i) SDE with $\eta = 0.75$, (i) ODE with $\eta = 0.75$; (iii) SDE with $\eta = 1.2$, (iv) ODE with $\eta = 1.2$.

F.3 DDPO Images under PickScore

Figure 8: Rows (from left to right): (i) SDE with $\eta = 0.75$, (i) ODE with $\eta = 0.75$; (iii) SDE with $\eta = 1.5$, (iv) ODE with $\eta = 1.5$.

F.4 DDPO Reward Gaps for other rewards

F.4.1 ImageReward

F.4.2 HPSv2

F.4.3 AESTHETIC