Value-as-Return: A Two-Stage Framework to Align on the Optimal Score Function

Shikun Sun [*1] **Shuo Huang** [*1] **Yiding Chen** [2] **Wen Sun** [2] **Jia Jia** [1 3]

Figure 1. Generated samples of VRPO-SDXL. See Appendix H for the corresponding prompts.

Abstract

Reinforcement learning with diffusion models has shown strong potential, but existing approaches such as variants of Direct Preference Optimization (DPO) often rely on an inaccurate simplification: they equate trajectory likelihoods with final-state probabilities. This mismatch leads to suboptimal alignment. We address this limitation with a principled framework that leverages the optimal value function as the return for short trajectory segments. Our approach follows a two-stage procedure: (i) learning a value-distribution function to estimate segment-level returns, and (ii) applying our VRPO to refine the score function. We prove that, under sufficient model capacity, the resulting model is equivalent to training a diffusion process on the tilted distribution proportional to $p(x)\exp(\eta r(x))$. Experiments on large-scale diffusion models validate our analysis and show stable and consistent improvements over prior methods. Code is available at https://github.com/huangshuo200823/VRPO.

[*]Equal contribution [1]Department of Computer Science and Technology, Tsinghua University [2]Cornell University [3]BNRist, Tsinghua University. Correspondence to: Jia Jia <jjia@tsinghua.edu.cn>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Reinforcement learning(Wallace et al., 2023; Black et al., 2023; Agarwal et al., 2019) (RL) has become an increasingly important tool for guiding generative models, and diffusion models (Song et al., 2021a;b; Ho et al., 2020) in particular, toward producing outputs that are both high-quality and well-aligned with task-specific objectives. By incorporating reward signals, RL provides a way to capture complex preferences that are not easily expressed through likelihood-based training alone. Despite these advantages, applying RL to diffusion models remains challenging due to the difficulty of faithfully integrating rewards defined in the image domain into the trajectory-based training process. A prominent example is given by existing variants of Direct Preference Optimization (Wallace et al., 2023; Rafailov et al., 2023) (DPO), which commonly treat pairwise preferences over final states as if they directly reflected trajectory-level preferences. This oversimplification often

misrepresents the underlying reward structure, leading to misaligned policies and degraded generation quality.

Several recent works (Yang et al., 2023; Liang et al., 2024) attempt to mitigate this issue by decomposing long diffusion trajectories into shorter segments. However, such approaches typically rely on heuristic design choices or assumptions that may not consistently hold in practice. As a result, they can introduce inconsistencies across time steps and undermine the stability of reward propagation, further complicating optimization and limiting generalization.

In this work, we propose a principled two-stage RL framework for diffusion models that is grounded in stochastic optimal control (Domingo-Enrich et al., 2025)(SOC). **The key idea is to leverage the optimal value function as the return signal for short trajectory segments**, thereby aligning training objectives more closely with the true reward structure. Our framework proceeds in two stages. In the first stage, we learn an optimal value-distribution function that characterizes the distribution of returns from intermediate states. In the second stage, we employ our stage-wise VRPO to refine the diffusion model, using preferences derived from the learned value function. This decomposition allows us to preserve reward consistency across segments while making long-horizon optimization tractable.

Our contributions are threefold:

- We introduce a two-stage RL framework that combines optimal value-distribution function training with our VRPO, providing a principled approach to align on the optimal score function.
- We prove that, under sufficient model capacity, our method converges to the optimal solution of training a diffusion process on the tilted distribution, which is proportional to $p(\mathbf{x})\exp(\eta r(\mathbf{x}))$, where r is the reward function in the image domain, and $p(\cdot)$ is the original distribution.
- We propose a CDF-based training strategy for the optimal value-distribution function, enhancing the training stability and robustness.

We validate our theoretical findings through extensive experiments on large-scale diffusion models(Rombach et al., 2022; Podell et al., 2023), demonstrating stable and consistent performance improvements across a variety of tasks.

2. Preliminary

2.1. Diffusion Generation Process as Continuous Time Markov Chain

To be clear, all the notions ignore the text condition c. Suppose the forward diffusion process is defined by the following SDE:

$$\mathrm{d}\mathbf{x}'_\tau = -\mathbf{f}(\mathbf{x}'_\tau, 1-\tau)\,\mathrm{d}\tau + \sigma(1-\tau)\,\mathrm{d}\mathbf{B}'_\tau, \qquad \tau \in [0,1], \tag{1}$$

where $\mathbf{x}'_\tau \in \mathbb{R}^d$ is the state, $\mathbf{B}'_\tau$ is a d-dimensional standard Brownian motion, $\mathbf{f} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is the drift, and $\sigma : [0,1] \to \mathbb{R}_+$ is the diffusion scale (e.g., VP/VE families). The process starts from $\mathbf{x}'_0 \sim p_{\text{data}}$ (data distribution), and ends at $\mathbf{x}'_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Then, the reverse-time SDE is given by (Anderson, 1982; Song et al., 2021b):

$$\begin{aligned}\mathrm{d}\mathbf{x}_t = &\Big(\mathbf{f}(\mathbf{x}_t, t) + \sigma(t)^2\,\nabla_{\mathbf{x}} \log p_{\text{ref}}(\mathbf{x}_t, t)\Big)\,\mathrm{d}t \\ &+ \sigma(t)\,\mathrm{d}\mathbf{B}_t, \qquad t \in [0,1].\end{aligned} \tag{2}$$

where $\mathbf{B}_t$ is a d-dimensional standard Brownian motion, $p_{\text{ref}}(\mathbf{x}, t)$ denotes the (reference/base) marginal density of $\mathbf{x}_t$ at time t. Note that $\mathbf{x}_t$ and $\mathbf{x}'_{1-t}$ follow the same distribution. We use boldface $\mathbf{x}$ for states and write $\boldsymbol{x}_{0:1} = \{\mathbf{x}_t : t \in [0,1]\}$ to denote a trajectory.

The reverse-time SDE (2) defines a continuous-time Markov chain (CTMC) whose transition law is fully determined by the score $\nabla_{\mathbf{x}} \log p_{\text{ref}}(\mathbf{x}, t)$ together with $\mathbf{f}$ and σ.

2.2. Formulation From Stochastic Optimal Control

We adopt the SOC objective with zero running cost and terminal cost $-\eta r(\mathbf{x})$ (terminal reward $r : \mathbb{R}^d \to \mathbb{R}$ with coefficient $\eta > 0$), and define the *control* as the score gap

$$\frac{1}{\sigma(t)}\mathbf{u}(\mathbf{x}, t) \;:=\; \mathbf{s}_\theta(\mathbf{x}, t) - \nabla_{\mathbf{x}} \log p_{\text{ref}}(\mathbf{x}, t), \tag{3}$$

so $\mathbf{u} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ has the same dimension as $\mathbf{x}$.

According to the definition in (Domingo-Enrich et al., 2025; Bellman, 1966), the controlled reverse-time SDE is

$$\mathrm{d}\mathbf{x}^{\mathbf{u}}_t = \big(\mathbf{b}_0(\mathbf{x}^{\mathbf{u}}_t, t) + \sigma(t)\,\mathbf{u}(\mathbf{x}^{\mathbf{u}}_t, t)\big)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\mathbf{B}_t, \tag{4}$$

where $\mathbf{b}_0(\mathbf{x}, t) := \mathbf{f}(\mathbf{x}, t) + \sigma(t)^2\,\nabla_{\mathbf{x}} \log p_{\text{ref}}(\mathbf{x}, t)$, and $\mathbf{x}^{\mathbf{u}}_t$ is the controlled process under $\mathbf{u}$. Let $p^{\mathbf{u}}(\boldsymbol{x} \mid \mathbf{x}_0)$ denote the path measure of the trajectory $\boldsymbol{x}_{0:1}$ induced by (4) given the initial state $\mathbf{x}_0$.

The optimal control problem is to find $\mathbf{u}^*$ minimizing the expected cost:

$$\min_{\mathbf{u}}\; \mathbb{E}_{\boldsymbol{x}^{\mathbf{u}} \sim p^{\mathbf{u}}}\left[\int_0^1 \tfrac{1}{2}\|\mathbf{u}(\mathbf{x}^{\mathbf{u}}_t, t)\|^2\,\mathrm{d}t - \eta r(\mathbf{x}^{\mathbf{u}}_1)\right], \tag{5}$$

where $\mathbf{x}^{\mathbf{u}}_1$ is the terminal state of the controlled trajectory. With the Theorem 2 in (Domingo-Enrich et al., 2025), we have

$$\begin{aligned}D_{\mathrm{KL}}\Big(p^{\mathbf{u}}(\boldsymbol{x} \mid \mathbf{x}_0)\,\Big\|\,p_{\text{ref}}(\boldsymbol{x} \mid \mathbf{x}_0)\Big) = \\ \mathbb{E}_{\boldsymbol{x}^{\mathbf{u}} \sim p^{\mathbf{u}}}\left[\int_0^1 \tfrac{1}{2}\|\mathbf{u}(\mathbf{x}^{\mathbf{u}}_t, t)\|^2\,\mathrm{d}t\right].\end{aligned} \tag{6}$$

which shows that the optimal control problem is equivalent to the KL-regularized continuous-time RL(Doya, 2000) problem:

$$\max_{\mathbf{u}\in\mathcal{U}} \mathbb{E}_{\mathbf{x}_0\sim p_0}\Big[\mathbb{E}_{\boldsymbol{x}\sim p^{\mathbf{u}}(\cdot|\mathbf{x}_0)}\eta\, r(\mathbf{x}_1^{\mathbf{u}}) - D_{\mathrm{KL}}\big(p^{\mathbf{u}}(\boldsymbol{x}\mid\mathbf{x}_0)\,\big\|\,p_{\text{ref}}(\boldsymbol{x}\mid\mathbf{x}_0)\big)\Big]. \tag{7}$$

where p_0 is the prior over $\mathbf{x}_0$ (e.g., $\mathcal{N}(\mathbf{0},\mathbf{I})$).

2.3. DPO Formulation of Diffusion Models: A Bandit View

In the bandit view, we regard the diffusion model as a policy that directly produces the final output $\mathbf{x}_1$ with terminal marginal distribution

$$\pi_\theta(\mathbf{x}_1),$$

i.e., π_θ denotes the distribution of the terminal state under the learned policy (not necessarily equal to p_{ref}).

With this interpretation, the DPO(Rafailov et al., 2023) loss is written directly in terms of the final-sample probabilities:

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\,\mathbb{E}_{(\mathbf{x}_1^+,\mathbf{x}_1^-)\sim\mathcal{D}}\Bigg[\log\mathrm{sigmoid}\left(\frac{1}{\eta}\log\frac{\pi_\theta(\mathbf{x}_1^+)}{\pi_{\mathrm{ref}}(\mathbf{x}_1^+)} - \frac{1}{\eta}\log\frac{\pi_\theta(\mathbf{x}_1^-)}{\pi_{\mathrm{ref}}(\mathbf{x}_1^-)}\right)\Bigg]. \tag{8}$$

where π_{ref} is the terminal marginal of the reference process.

A widely held (but over-simplified) assumption. We let the discrete time grid be $\mathcal{T}_T := \{0, \frac{1}{T}, \dots, 1\}$, and adopt the reverse-time generative process that starts from $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and evolves to the terminal time $\mathbf{x}_1$. It is common in practice to *identify* the marginal likelihood of the final sample with the likelihood of a *single realized trajectory*, i.e.,

$$\pi_\theta(\mathbf{x}_1) \overset{\text{(over-simplified)}}{\approx} p_{\text{ref}}(\mathbf{x}_0)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{\frac{t}{T}} \mid \mathbf{x}_{\frac{t-1}{T}}), \tag{$\star$}$$

where the right-hand side is evaluated along one generated path $\mathbf{x}_0 \to \cdots \to \mathbf{x}_1$. However, the right-hand side of $(\star)$ is the *joint* density of a full trajectory $\boldsymbol{x}_{0:1}$, not the *marginal* density of $\mathbf{x}_1$. This notation conflates a path-wise likelihood with the terminal marginal, and underlies many heuristic DPO adaptations.

The simplification in $(\star)$ obscures two practical issues:

- **Intractability of the marginal.** The true $\pi_\theta(\mathbf{x}_1)$ requires marginalizing over *all* trajectories that lead to $\mathbf{x}_1$:

$$\pi_\theta(\mathbf{x}_1) = \int\cdots\int p_{\text{ref}}(\mathbf{x}_0)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{\frac{t}{T}} \mid \mathbf{x}_{\frac{t-1}{T}})\,\mathrm{d}\mathbf{x}_{\frac{1}{T}:\frac{T-1}{T}}. \tag{9}$$

  Replacing this integral by a single path likelihood introduces bias.

- **Memory/compute burden.** The horizon T is large(Yang et al., 2023; Liang et al., 2024); storing all intermediate states and their gradients to approximate the marginal (or to backpropagate through all steps) is often impractical in GPU memory.

3. Problems in Current Diffusion DPO Methods

Therefore, there are mainly three problems in current diffusion DPO methods:

- **Inconsistency between trajectory-level probability and final state probability.** Current methods (Wallace et al., 2023; Liang et al., 2024; Yang et al., 2023) assume that the trajectory-level probability can be approximated by the final state probability, which is not always true. This assumption can lead to suboptimal policies that do not align well with the true reward function.
- **Inefficiency in utilizing reward signals.** Current methods do not fully leverage the reward signals available at each time step, or not pay attention to the consistency of rewards at different time steps, leading to inefficient learning and suboptimal performance.
- **Lack of theoretical guarantees.** Many existing approaches (Yang et al., 2023; Liang et al., 2024) lack rigorous theoretical analysis, making it difficult to understand their convergence properties and optimality. This can lead to uncertainty about the reliability and robustness of the learned policies, especially compared with original DPO (Rafailov et al., 2023) in NLP.

4. Methodology

To address the limitations of existing diffusion DPO methods, we introduce a two-stage framework that leverages the optimal value function as the return for short trajectory segments. This decomposition enables long trajectories to be broken into manageable pieces, where preferences can be assessed more reliably. We provide a detailed roadmap of our methods in Appendix J.

4.1. Overall Pipeline

Our pipeline consists of two stages. In the first stage, we train the optimal value-distribution function. In the sec-

ond stage, we perform Value-as-Return Direct Preference Optimization (VRPO) using this learned value function.

The framework is built on two key insights. First, **when trajectories are sufficiently short, the preference over the final state coincides with the preference over the entire trajectory, conditioned on the same starting point.** Moreover, the memory burden is also alleviated. This observation justifies segmenting long trajectories into shorter ones, on which DPO can be applied directly. Second, **the optimal value function admits a closed-form expression that depends only on the reference policy and the terminal reward in the image domain.** As a result, we can learn a value-distribution function to estimate segment-level preferences, and then employ these preferences to guide VRPO training.

4.2. First Stage: Training the Optimal Value-Distribution Function

Value-distribution definition (terminal at $t = 1$). We define the *optimal value-distribution function* through the random variable

$$Z(\mathbf{x}_t, t) \;:=\; r(\mathbf{x}_1), \qquad \mathbf{x}_1 \sim p_{\text{ref}}(\mathbf{x}_1 \mid \mathbf{x}_t, t), \quad t \in \mathcal{T}_T, \tag{10}$$

where $p_{\text{ref}}(\mathbf{x}_1 \mid \mathbf{x}_t, t)$ is the conditional distribution of the reference diffusion model from intermediate state $(\mathbf{x}_t, t)$ to the terminal state $(\mathbf{x}_1, 1)$, and $r(\cdot)$ is the image-domain reward defined at $t = 1$. Then, we directly have the following closed-form expression for the optimal value function:

Proposition 1. *The optimal value function of Equation* (7) *admits the closed-form*

$$\begin{aligned} V^*(\mathbf{x}_t, t, \eta) &= \log\, \mathbb{E}\big[\exp\!\big(\eta\, Z(\mathbf{x}_t, t)\big)\big] \\ &= \log\, \mathbb{E}_{\mathbf{x}_1 \sim p_{ref}(\mathbf{x}_1 \mid \mathbf{x}_t, t)}\big[\exp\!\big(\eta\, r(\mathbf{x}_1)\big)\big]\,. \end{aligned} \tag{11}$$

which depends only on the reference policy p_{ref} and the terminal reward r. Consequently, once the distribution $Z(\mathbf{x}_t, t)$ is learned, the KL coefficient $\frac{1}{\eta}$ can be adjusted post hoc *without retraining.*

We explored multiple approaches for modeling the distribution of $Z(\mathbf{x}_t, t)$, and observed that training with a mix of the Continuous Ranked Probability Score (CRPS) loss, Binary Cross Entropy loss and smooth loss to fit the cumulative distribution function (CDF) yields more stable performance than directly fitting the probability density function (PDF) using maximum likelihood estimation (MLE) or KL divergence, as shown in Algorithm 2. Accordingly, we parameterize a model $F_\phi(z \mid \mathbf{x}_t, t)$ to approximate the CDF of $Z(\mathbf{x}_t, t)$, and compute the optimal value function for different η as

$$V^*_\phi(\mathbf{x}_t, t, \eta) \;=\; \log \int e^{\eta z}\, dF_\phi(z \mid \mathbf{x}_t, t). \tag{12}$$

4.3. Second Stage: Stage-wise VRPO Training

We decompose the long trajectory on $\mathcal{T}_T$ into short segments of length L steps, so that $t_k \in \{0, \frac{1}{T}, \ldots, 1 - \frac{L}{T}\}$ and the segment runs from t_k to $t_k + \frac{L}{T}$. Under the short-segment assumption, *the terminal-state preference within the segment aligns with the trajectory preference given the same start*. We therefore evaluate segment preferences using the optimal value function at the segment endpoint. The overall procedure is shown in Algorithm 1.

Algorithm 1 Stage-wise VRPO with Learned Value V^*_ϕ (time grid $0, \frac{1}{T}, \ldots, 1$)

Require: Reference transition $p_{\text{ref}}(\mathbf{x}_{t+\frac{1}{T}} \mid \mathbf{x}_t)$; learned $q_\phi(z \mid \mathbf{x}_t, t)$; model p_θ to train; segment length L; #segments K; DPO temperature $\beta = \frac{1}{\eta}$

Value from the learned distribution:

$$V^*_\phi(\mathbf{x}_t, t, 1/\beta) \;=\; \log \mathbb{E}_{z \sim q_\phi(\cdot \mid \mathbf{x}_t, t)}\big[\exp\big((1/\beta)\, z\big)\big].$$

// Stage-wise over short segments

for each training iteration **do**

  Sample start time $t_k \in \{0, \frac{1}{T}, \ldots, 1 - \frac{L}{T}\}$ and start state $\mathbf{x}_{t_k} \sim p_{\text{ref}}(\mathbf{x}_{t_k}, t_k)$

  From $\mathbf{x}_{t_k}$, **roll out** L **steps** with the *reference* model to the endpoint $\mathbf{x}_{t_k + \frac{L}{T}}$

  Generate candidate endpoints (e.g., two independent rollouts) $\mathbf{x}^+_{t_k + \frac{L}{T}}$ and $\mathbf{x}^-_{t_k + \frac{L}{T}}$ and **rank** by

$$V^*_\phi(\mathbf{x}^+_{t_k + \frac{L}{T}},\, t_k + \tfrac{L}{T},\, \tfrac{1}{\beta}) \;>\; V^*_\phi(\mathbf{x}^-_{t_k + \frac{L}{T}},\, t_k + \tfrac{L}{T},\, \tfrac{1}{\beta}).$$

  Update p_θ with the **VRPO** loss:

$$\begin{aligned} \mathcal{L}_{\text{VRPO}} = -\mathbb{E}\,\Big[\log \text{sigmoid}\Big(&\beta \log p_\theta\big(\mathbf{x}^+_{t_k + \frac{L}{T}} \mid \mathbf{x}_{t_k}\big) \\ &-\beta \log p_{\text{ref}}\big(\mathbf{x}^+_{t_k + \frac{L}{T}} \mid \mathbf{x}_{t_k}\big) - \beta \log p_\theta\big(\mathbf{x}^-_{t_k + \frac{L}{T}} \mid \mathbf{x}_{t_k}\big) \\ &+\beta \log p_{\text{ref}}\big(\mathbf{x}^-_{t_k + \frac{L}{T}} \mid \mathbf{x}_{t_k}\big)\Big)\Big]. \end{aligned}$$

end for

Output: A VRPO–trained diffusion model p_θ.

Proposition 2. *Suppose $\frac{L}{T} = \Delta$ is sufficiently small, then we have that given the same start $(\mathbf{x}_{t_k}, t_k)$, the endpoint probability can be approximated by a one-step (Euler–Maruyama) approximation:*

$$p(\mathbf{x}_{t_k + \frac{L}{T}} \mid \mathbf{x}_{t_k}, t_k) \approx \mathcal{N}\big(\, \mu_{t_k},\; \Sigma_{t_k}\,\big), \tag{13}$$

where

$$\begin{aligned} \mu_{t_k} &= \mathbf{x}_{t_k} + \big(f(\mathbf{x}_{t_k}, t_k) + \sigma(t_k)^2\, s_\theta(\mathbf{x}_{t_k}, t_k)\big)\,\Delta, \\ \Sigma_{t_k} &= \sigma(t_k)^2\, \Delta\, \mathbf{I}. \end{aligned} \tag{14}$$

We use Proposition 2 to ensure that the endpoint distribution of a short trajectory segment can be well approximated. In

practice, however, one may also work with the trajectory probability because of the short trajectory. Then, we have:

Theorem 3 (Optimality of VRPO)**.** *Consider a trajectory on $\mathcal{T}_T$ that is decomposed into K short segments, each of length L with $\Delta = \frac{L}{T}$ sufficiently small. If VRPO is applied to segment-level preferences defined by $V_k^*(\mathbf{x}_t, t, \eta)$, then under sufficient model capacity, the resulting model is equivalent to training directly on the optimal target distribution*

$$p^*(\mathbf{x}_1) \;\propto\; p_{data}(\mathbf{x}_1)\, \exp\!\big(\eta\, r(\mathbf{x}_1)\big). \tag{15}$$

Remark 4 (Well-definedness and Uniqueness)**.** *Theorem 3 does not merely characterize the optimal terminal distribution at $t = 1$. Rather, it establishes a **pathwise optimal solution** that is unique: the learned optimal score function $s_\theta^*(\mathbf{x}_t, t)$ coincides, for all $t \in [0, 1]$, with the score of the time-marginal distribution induced by the optimal controlled process. Importantly, this property is a necessary condition for switching between SDE-based and ODE-based sampling procedures without training an additional network.*

5. Experiments

5.1. Implementation Details

5.1.1. Value Function Training

For the CDF model $F_\phi(z \mid \mathbf{x}_t, t)$, we adopt the PickScore v1 model (Kirstain et al., 2023) as our backbone. PickScore builds on the CLIP (Radford et al., 2021) dual-encoder architecture, where a text encoder and an image encoder map prompts and images into a shared embedding space, and their dot product (scaled by a learnable parameter) serves as the preference score. The model is fine-tuned on human preference data using a pairwise ranking loss, aligning the scores with users' comparative judgments of image quality. In our framework, the parameters of the PickScore encoders are kept frozen, so that the extracted text and image embeddings remain fixed. On top of these embeddings, we design a lightweight prediction head: a timestep embedding is first combined with the image embedding, and the resulting representation is fused with the text embedding through a MLP. Finally, the fused features are passed into a monotonic CDF (cumulative distribution function) Head, which enforces non-decreasing outputs along discretized thresholds and produces a valid CDF output.

We train the model following Algorithm 2. Instead of rolling out a trajectory at each training step, we construct a dataset of images $\mathbf{x}_1$ using the prompts from the training split of the Pic-a-Pic v1 dataset (Kirstain et al., 2023). During training, these images are further perturbed with DDIM (Song et al., 2021a) to obtain $\mathbf{x}_t$, which are then used as inputs. The PickScore value serves as the ground-truth reward $r(\mathbf{x}_1)$. By default, we set the horizon to $T = 20$, and the number of discrete classes to $M = 50$. For the SD-1.5 (Rombach et al., 2022) experiments, PickScore values are uniformly discretized into 50 bins over the range [15, 25], while for SDXL (Podell et al., 2023) the range [17, 27] is divided into 50 bins. Training is performed with the AdamW (Loshchilov & Hutter, 2017) optimizer, using a learning rate of 1×10^{-4}, betas $(0.9, 0.98)$, $\epsilon = 1 \times 10^{-6}$, and a weight decay of 0.01. We adopt a learning rate schedule with warm-up followed by cosine annealing. The overall loss is defined as a weighted sum of three components: a CRPS loss with weight 0.7, a cross-entropy loss with weight 0.3, and a first-order total variation smoothness loss.

5.1.2. VRPO Training

We train the VRPO on the training prompts of the Pick-a-Pic dataset. We employ DDIM sampling with 20 denoising steps, combined with classifier-free guidance at a scale of 5.0. Both SD-1.5 and SDXL are fine-tuned using LoRA (Hu et al., 2021) adapters for 15 epochs, with the LoRA rank set to 4 for SD-1.5 and 64 for SDXL. A regularization coefficient of $\beta = 10$ (with $\eta = \frac{1}{\beta} = 0.1$) is applied throughout training. For SD-1.5, we adopt a batch size of 40 and a learning rate of 2×10^{-5}, while for SDXL we use a batch size of 8 with gradient accumulation over 2 steps and a learning rate of 1.5×10^{-5}. Optimization is performed with AdamW, configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1×10^{-4}. The number of inner optimization steps L is set to 1 for SD-1.5 and 4 for SDXL.

5.2. Comparing with Baselines

5.2.1. Quantitative Comparison

Problem Settings. We compare our method VRPO against a simple baseline: using our value function as classifier guidance, and three strong baselines: MAPO (She et al., 2024), Diffusion-DPO (Wallace et al., 2023), and SPO (Liang et al., 2024). MAPO and Diffusion-DPO are trained on the Pick-a-Pic v2 dataset, which contains over 800,000 image–preference pairs, whereas SPO and our method use the Pick-a-Pic v1 dataset with about 500,000 pairs. For fairness, we adopt the official checkpoints of all baselines. Our evaluations are performed on the "validation_unique" split of Pick-a-Pic v1: after filtering out prompts with special characters, we obtain 492 valid prompts.

We adopt four evaluation metrics: PickScore v1, ImageReward (Xu et al., 2023), Aesthetic Score (Schuhmann, 2022), and HPSv2 (Wu et al., 2023). PickScore v1 is a CLIP-based preference predictor trained on human data, and has been shown to correlate more strongly with human rankings than earlier automated metrics. HPSv2 is a preference prediction model trained on the large-scale HPD v2 dataset, designed to better capture human judgments across diverse image distributions. Aesthetic Score evaluates the visual appeal of generated images in a prompt-agnostic manner, while Im-

Table 1. Results of quantitative comparison with baselines and ablation studies on SDXL.

METHODS	PICKSCORE	IMAGEREWARD	AESTHETIC SCORE	HPSV2
SDXL	21.7093	0.5887	6.4658	0.2738
CLASSIFIER GUIDANCE	21.7768	0.6022	6.4716	0.2738
MAPO	21.9038	0.7429	6.5918	0.2791
DIFFUSIONDPO	22.3269	0.9411	6.3959	0.2837
SPO	22.7821	1.0536	6.8035	0.2858
VRPO (OURS)	**23.0656**	**1.2404**	**7.1894**	**0.2867**
FULL METHODS	23.0656	**1.2404**	**7.1894**	**0.2867**
W/O OVF	22.9077	1.1439	6.8553	0.2825
CE LOSS FOR LOGITS	22.6795	1.0325	6.7490	0.2832
AVERAGE SCORES	22.9472	1.1308	6.8737	0.2843
BINS WIDTH 20	22.4321	1.0172	6.7177	0.2825
BINS WIDTH 30	22.8641	1.1605	6.7899	0.2823
BINS WIDTH 75	23.0070	1.2186	7.1854	0.2839
BINS WIDTH 100	**23.0923**	1.2008	7.0816	0.2859
W/O OVF + TRAJECTORY	22.6980	1.0378	6.7569	0.2829

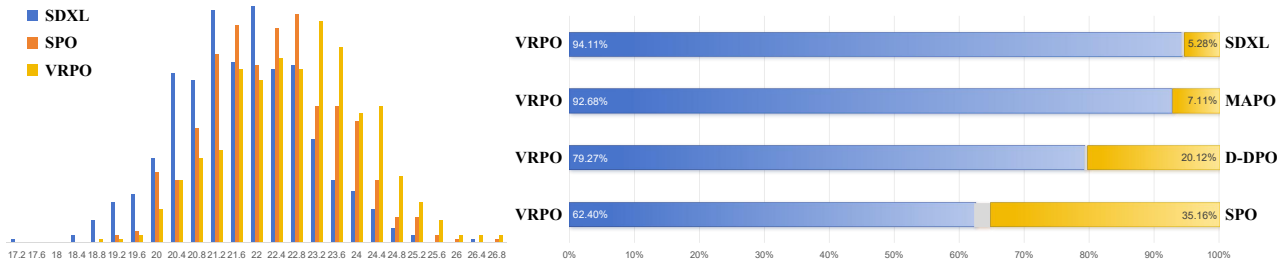

Figure 2. Left: Histograms of PickScore values for SDXL, SPO, and VRPO. *Right*: Win rates of VRPO compared to other baselines in terms of PickScore values.

ageReward aggregates multiple quality features to provide a holistic assessment of image quality.

Quantitative comparison. The results of the quantitative comparison are reported in Table 1 and Table 2. To provide a more intuitive understanding of the superiority of VRPO, we further present the histogram of PickScore values and the win rate comparison in Figure 2. When applied to SD v1.5, our method outperforms all other methods across all metrics by a substantial margin. For SDXL, our method demonstrates clear advantages in PickScore, ImageReward, and Aesthetic Score. Notably, in the supervised metric of PickScore, we outperform the baseline by more than 0.65 on SD v1.5 and 0.19 on SDXL. When comparing with the best-performing baseline, the SPO method, our approach uses significantly fewer parameters for training the value function but achieves superior results. This further supports the our effectiveness of utilizing reward signals. Specifically, SPO incorporates the timestep into the input of the CLIP image encoder and fine-tunes the entire CLIP model. In contrast, we freeze the CLIP encoder, and finetune a light-weight head. Remarkably, while delivering state-of-the-art performance, VRPO also consumes the least training time across all baselines (see detailed comparison in Appendix F).

5.2.2. QUALITATIVE COMPARISON.

Figure 3 presents qualitative comparisons across different methods. All images are generated with the same random seed to enable a direct visual comparison of the effects of different fine-tuning strategies. Compared with the baselines, VRPO produces images with richer details and more coherent compositions.

5.2.3. HUMAN PREFERENCE COMPARISON.

In this study, we employed ChatGPT-5.0 as an evaluator to assess the quality of generated images by interpreting and articulating human-like judgments. Following the experimental setup of SPO, we randomly sampled 100 prompts from Parti-Prompts (Yu et al., 2022) and 200 prompts from

Figure 3. Comparison with baseline methods. See Appendix H for the corresponding prompts.

Table 2. Results of quantitative comparison with baselines and ablation studies on SD-1.5.

METHODS	PICKSCORE	IMAGEREWARD	AESTHETIC SCORE	HPSV2
SD V1.5	20.346	-0.089	5.956	0.235
SPO	21.1982	0.2726	6.4497	0.2712
OURS	**21.7609**	**0.6704**	**6.5209**	**0.2803**
W/O OVF	21.7183	0.5544	6.4832	0.2770
W/O OVF + TRAJECTORY	21.5579	0.4285	6.4681	0.2733

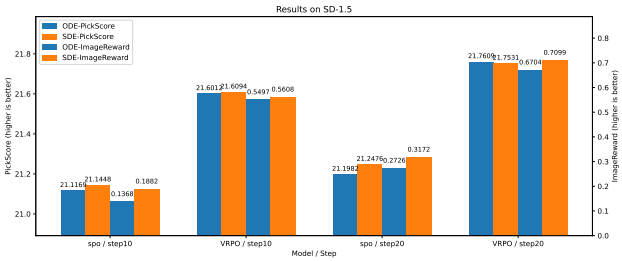

Figure 4. PickScore and ImageReward under different sampling strategies and differnt samping steps for SPO and VRPO.

HPSv2 (Wu et al., 2023), both unseen by VRPO and SPO, covering a broad range of categories. GPT-5.0 was then instructed to compare the outputs of SPO and our method along three dimensions: overall preference, visual appeal, and text–image alignment. More details are provided in Appendix I.

The results shown in Table 3 indicate that our method consistently received more "winning votes" from users in both general preference and visual appeal. Notably, the margin by which our method outperforms SPO in terms of visual appeal is particularly substantial. Please note that we also provided objective evaluation metrics in Table 4 in Appendix. These metrics revealed that our method demonstrated even more pronounced advantages compared to the results obtained from the Pick-a-Pic dataset. We additionally conduct a human user study, as reported in Appendix K.

5.2.4. SAMPLING ALGORITHM ANALYSIS.

To validate Remark 4, we compare the PickScore and ImageReward of SPO and VRPO, as shown in Figure 4. When switching the sampling procedure from the SDE-based rollout used during RL training to the corresponding ODE sampler, VRPO exhibits a noticeably smaller performance drop than SPO. This effect is particularly pronounced for ImageReward under 10-step sampling (-0.111 vs -0.514), and for PickScore under 20-step sampling (+0.0078 vs -0.0494).

Table 3. Preference comparison of VRPO against baseline methods (SDXL and SPO) in terms of overall preference, visual quality, and text–image alignment (reported as Win / Draw / Lose).

METHODS	OVERALL PERFORMANCE	VISUAL QUALITY	TEXT ALIGNMENT
VRPO V.S SDXL	**80.6**% / 0% / 19.4%	**76.8**% / 4.1% / 19.1%	**55.4**% / 17.6 % / 27.0%
VRPO V.S SPO	**65.8**% / 0% / 34.2%	**50.2**% / 26.4% / 23.4%	**40.0**% / 22.5% / 37.5%

5.3. Ablation Study

We conduct a series of ablation experiments to validate the effectiveness of our two key design choices—(i) training the value function with rollout trajectories rather than unrelated image datasets (as in SPO), and (ii) adapting the optimal value function (OVF). Results are reported on both SD-1.5 and SDXL (see Table 2 and Table 1), with more detailed ablations provided on SDXL.

Impact of probabilistic distribution. We remove the optimal value function (*w/o OVF*) completely and train a regression-based value function directly on the rollout trajectory dataset, we observe consistent performance degradation across all evaluation metrics on both SD-1.5 and SDXL. This confirms that modeling the inherent uncertainty in diffusion sampling through a probabilistic value function is crucial for improved performance.

Loss formulation. To further examine the effect of the training objective, we replace the CRPS loss with only the cross-entropy loss and substitute the CDF Head with a simple probability head (*CE Loss for Logits*). This modification results in a significant performance drop, indicating that cross-entropy loss alone fails to provide stable convergence for the value function.

Effectiveness of optimal value function. Instead of using the optimal score function derived in Proposition 1, we also experiment with computing the average score of the predicted distribution (*Average Scores*). This approximation again results in degraded performance, demonstrating that the optimal score function enhances the ability of DPO methods to effectively leverage the reward.

Effect of bin width. We study the discretization granularity of the distribution by varying the number of bins (*Bin Width*). Given that PickScore values span a range of 10 for both SD-1.5 ([15, 25]) and SDXL ([17, 27]), reducing the number of bins to 30 causes a severe performance decline, whereas increasing the bin resolution beyond 50 yields no further noticeable improvements. This suggests that around 50 bins strike a good balance between expressiveness and efficiency.

Joint removal of contributions. Finally, we remove both the roll-out-trajectory dataset and the probabilistic formulation (*w/o OVF + Trajectory*), which leads to a drastic performance drop. Comparing this result with *w/o OVF*, we observe that the dataset alone also contributes substantially to the overall performance.

More ablation studies on segment length and miss-ranking rate are provided on Appendix G.

6. Discussion

The proposed framework is not tied to a specific optimization scheme. First, the reward can be directly incorporated into GRPO (Shao et al., 2024) and other preference-based objectives, and it naturally extends to classical policy-gradient methods such as REINFORCE (Williams, 1992), where it provides a step-wise reward signal that helps reduce variance. Second, by decomposing a fixed-horizon MDP into a sequence of bandit-style subproblems, the framework admits a well-defined and unique solution at each step, which in turn preserves the structural simplicity of the induced diffusion process. Together, these properties suggest that our framework is broadly compatible with a wide family of reinforcement-learning algorithms.

A further distinction arises when comparing reinforcement learning in diffusion models to reinforcement learning in natural language processing. In NLP, policies are discrete and autoregressive, which makes trajectory probabilities tractable. Diffusion models, by contrast, evolve under continuous-time stochastic dynamics, where the marginal distribution of the final state requires integrating over all possible paths. Our framework addresses this gap by reformulating the problem in terms of value functions, thereby avoiding intractable marginalization while maintaining theoretical alignment with the target tilted distribution.

7. Conclusion

We presented a two-stage reinforcement learning framework for diffusion models that leverages the optimal value function as the return signal for short trajectory segments. By first learning a value-distribution function and then applying stage-wise VRPO, our method provides a principled approach to align diffusion models with image-domain rewards. We established theoretical guarantees showing that, under sufficient model capacity, the resulting model is equivalent to training on the tilted distribution proportional to $p(\mathbf{x})\exp(\eta r(\mathbf{x}))$. Extensive empirical results confirm the effectiveness of our approach, yielding stable and consistent improvements over prior online DPO-based methods.

Impact Statement

This paper studies methods for preference optimization in diffusion models. Our contributions are mainly methodological. Potential positive impacts include improving controllability of generative models, which may benefit applications such as art creation and scientific visualization. However, misuse of the method could lead to harmful content generation or reinforce biases present in the training data. We rely only on publicly available datasets in our experiments, and we encourage future work to carefully consider fairness, safety, and data governance issues when deploying such models.

Acknowledgements

This work is supported by the National Natural Science Foundation of China Nos. 62425604. It is also supported by Beijing Natural Science Foundation (L257006), High Performance Computing Center, Tsinghua University.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Bellman, R. Dynamic programming. *science*, 153(3731): 34–37, 1966.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Domingo-Enrich, C., Drozdzal, M., Karrer, B., and Chen, R. T. Q. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *ICLR*, 2025.

Doya, K. Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245, 2000.

Fleming, W. H. and Soner, H. M. *Controlled Markov processes and viscosity solutions*. Springer, 2006.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kappen, H. J. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011, 2005.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

Lee, K., Li, X., Wang, Q., He, J., Ke, J., Yang, M.-H., Essa, I., Shin, J., Yang, F., and Li, Y. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18465–18475, 2025.

Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., and Zheng, L. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. *arXiv preprint arXiv:2406.04314*, 2024.

Liu, X., Wei, Y., Liu, M., Lin, X., Ren, P., Xie, X., and Zuo, W. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Oksendal, B. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

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. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042.

Schuhmann, C. Laion-aesthetics, 2022. Accessed: 2023 - 11- 10.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

She, S., Zou, W., Huang, S., Zhu, W., Liu, X., Geng, X., and Chen, J. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*, 2024.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Sun, Y., Liu, Y., Tang, Y., Pei, W., and Chen, K. Anycontrol: create your artwork with versatile control on text-to-image generation. In *European Conference on Computer Vision*, pp. 92–109. Springer, 2024.

Todorov, E. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

Todorov, E. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28): 11478–11483, 2009.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

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

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Li, Q., Shen, W., Zhu, X., and Li, X. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y. Scaling autoregressive models for content-rich text-to-image generation, 2022.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

A. Use of LLMs

We used large language models (LLMs) solely for grammar and spelling checking in the preparation of this paper. No LLMs were involved in generating research ideas, conducting experiments, analyzing results, or writing technical content. All scientific contributions and claims are entirely the work of the authors.

B. Related Work

Diffusion models. Diffusion probabilistic models (Ho et al., 2020; Song et al., 2021b;a) have emerged as state-of-the-art generative models across image, audio, and video domains. Their success largely stems from stable likelihood training and scalable architectures, with large-scale variants such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2023) demonstrating impressive generative capabilities. Recent advances (Ho & Salimans, 2022; Meng et al., 2021; Zhang et al., 2023; Nichol et al., 2021; Lee et al., 2025; Liu et al., 2024; Sun et al., 2024) have sought to improve controllability by conditioning on textual prompts, sketches, or layout, but aligning with more complex human-defined reward signals remains a challenge.

Reinforcement learning for generative models. RL has been widely adopted in language modeling (Rafailov et al., 2023; Ouyang et al., 2022), where it enables alignment with human preferences through pairwise comparisons. In diffusion models, early efforts adapted policy gradient (Agarwal et al., 2019; Black et al., 2023), but these methods often suffer from high variance and unstable credit assignment across long horizons. Recent works therefore explore preference-based training tailored to diffusion models, such as Diffusion-DPO (Wallace et al., 2023) and D3PO (Yang et al., 2023), which directly optimize model likelihood ratios against a reference. However, these approaches typically equate trajectory likelihoods with terminal-state probabilities, an oversimplification that can misalign the learned distribution with the intended reward structure.

Segment-wise training. To mitigate horizon-related difficulties, several works decompose long diffusion rollouts into shorter segments (Liang et al., 2024; Yang et al., 2023). While this improves stability, segment-level objectives are often designed heuristically, and their consistency with the global reward remains unclear. Our work provides a principled alternative: we ground the segment-wise objective in stochastic optimal control, derive a closed-form expression of the optimal value function, and use it to define preference-consistent objectives for Direct Preference Optimization. This yields both theoretical guarantees and practical improvements.

Stochastic optimal control and continuous-time RL. Our approach is also connected to the literature on SOC and continuous-time RL. Linearly-solvable MDPs (Todorov, 2006; 2009) show that optimal controls can often be expressed via exponentiated value functions and Doob h-transforms, a perspective that closely parallels our formulation. In continuous-time settings, stochastic control theory provides tools such as the Hamilton–Jacobi–Bellman equation and Girsanov's theorem (Fleming & Soner, 2006; Oksendal, 2013), which describe how optimal policies reshape diffusion dynamics. Recent works (Domingo-Enrich et al., 2025) have further connected RL objectives with score-based generative modeling. Our method builds on these insights by importing SOC principles into diffusion preference optimization, establishing both theoretical grounding and practical benefits.

C. The Traning Algorithm of Value-Distribution Function

We have a pratical training algorithm as shown in Algorithm 2.

D. Key Analysis Results from SOC

D.1. Optimal Value Function

From the SOC formulation, the *value function* is defined as the optimal objective under the controlled reverse-time SDE (4):

$$V(\mathbf{x}, t) \;:=\; \min_{\mathbf{u}} \; \mathbb{E}_{\boldsymbol{x}^{\mathbf{u}} \sim p^{\mathbf{u}}} \left[\int_t^1 \tfrac{1}{2} \|\mathbf{u}(\mathbf{x}_s^{\mathbf{u}}, s)\|^2 \, \mathrm{d}s - \eta r(\mathbf{x}_1^{\mathbf{u}}) \,\middle|\, \mathbf{x}_t^{\mathbf{u}} = \mathbf{x} \right]. \tag{16}$$

Algorithm 2 Learning the Value-Distribution via CDF Modeling (CRPS + BCE + TV-Smooth)

Require: Reference sampler $p_{\text{ref}}(\mathbf{x}_1)$; reward $r(\mathbf{x}_1)$; forward noising $\text{Noisify}(\mathbf{x}_1, t)$; CDF model $F_\phi(z \mid \mathbf{x}_t, t)$; grid $\{c_m\}_{m=1}^M$; weights $(\lambda_{\text{crps}}, \lambda_{\text{bce}}, \lambda_{\text{tv}})$; small $\varepsilon > 0$

for each iteration **do**

Sample terminal images: $\{\mathbf{x}_1^{(i)}\}_{i=1}^B \sim p_{\text{ref}}(\mathbf{x}_1)$

Score images: $y^{(i)} \leftarrow r(\mathbf{x}_1^{(i)})$

Sample time and noising: $t^{(i)} \sim \text{Unif}(\mathcal{T}_T)$, $\mathbf{x}_{t^{(i)}}^{(i)} \leftarrow \text{Noisify}\big(\mathbf{x}_1^{(i)}, t^{(i)}\big)$

Evaluate CDF on grid: $F_m^{(i)} \leftarrow F_\phi\big(c_m \mid \mathbf{x}_{t^{(i)}}^{(i)}, t^{(i)}\big)$, $m = 1{:}M$

CRPS loss:

$$\mathcal{L}_{\text{crps}} \;=\; \frac{1}{B}\sum_{i=1}^{B}\sum_{m=1}^{M}\Big(F_m^{(i)} - \mathbf{1}\{c_m \geq y^{(i)}\}\Big)^2$$

CDF binary cross-entropy:

$$\mathcal{L}_{\text{bce}} \;=\; -\frac{1}{B}\sum_{i=1}^{B}\sum_{m=1}^{M}\Big[\mathbf{1}\{c_m \geq y^{(i)}\}\,\log(F_m^{(i)} + \varepsilon) + \big(1 - \mathbf{1}\{c_m \geq y^{(i)}\}\big)\,\log(1 - F_m^{(i)} + \varepsilon)\Big]$$

TV smoothness on CDF:

$$\mathcal{L}_{\text{tv}} \;=\; \frac{1}{B}\sum_{i=1}^{B}\frac{1}{M-1}\sum_{m=1}^{M-1}\Big|F_{m+1}^{(i)} - F_m^{(i)}\Big|$$

Total loss and update:

$$\mathcal{L} \;=\; \lambda_{\text{crps}}\mathcal{L}_{\text{crps}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}, \qquad \phi \leftarrow \phi - \eta\nabla_\phi\mathcal{L}$$

end for

Output: Trained CDF $F_\phi(z \mid \mathbf{x}_t, t)$.

Post hoc value:

$$V_\phi^*(\mathbf{x}_t, t, \eta) \;\approx\; \log\sum_{m=1}^{M}\tilde{\pi}_m(\mathbf{x}_t, t)\, e^{\eta c_m},$$

$$\tilde{\pi}_m(\mathbf{x}_t, t) = \frac{\big(F_\phi(c_m \mid \mathbf{x}_t, t) - F_\phi(c_{m-1} \mid \mathbf{x}_t, t)\big)_+ + \varepsilon}{\sum_{\ell=1}^{M}\big(F_\phi(c_\ell \mid \mathbf{x}_t, t) - F_\phi(c_{\ell-1} \mid \mathbf{x}_t, t)\big)_+ + M\varepsilon}.$$

Equivalently, by the linearly-solvable control result (Kappen, 2005), the value function admits a representation with respect to the uncontrolled reference process p_{ref}:

$$V(\mathbf{x}, t) \;=\; -\log\, \mathbb{E}_{\boldsymbol{x}\sim p_{\text{ref}}}\Big[\exp\big(\eta r(\mathbf{x}_1)\big)\,\Big|\,\mathbf{x}_t = \mathbf{x}\Big]\,, \tag{17}$$

where the expectation is over trajectories $\boldsymbol{x}_{0:1}$ of the base process.

D.2. Optimal Score Function

The optimal control $\mathbf{u}^*(\mathbf{x}, t)$ can be expressed via the gradient of the value function (Kappen, 2005):

$$\mathbf{u}^*(\mathbf{x}, t) \;=\; -\sigma(t)\,\nabla_{\mathbf{x}} V(\mathbf{x}, t). \tag{18}$$

Then the *optimal score function* is

$$\mathbf{s}^*(\mathbf{x}, t) \;=\; \nabla_{\mathbf{x}} \log p_{\text{ref}}(\mathbf{x}, t) - \nabla_{\mathbf{x}} V(\mathbf{x}, t). \tag{19}$$

It is worth noting that $\mathbf{s}^*(\mathbf{x}, t)$ corresponds to the score of the tilted distribution

$$p^*(\mathbf{x}) = \frac{1}{Z}\, p_{\text{ref}}(\mathbf{x})\, \exp(\eta r(\mathbf{x})),$$

where Z is the normalization constant.

E. Proofs

E.1. Proof of Lemma 1

Proof. Fix $(\mathbf{x}_t, t)$ and let $\mathbb{P}^{\text{ref}}_{\mathbf{x}_t,t}$ denote the (path) law on trajectories $(\mathbf{x}_s)_{s\in[t,1]}$ induced by the (reference) reverse-time SDE starting at $\mathbf{x}_t$ at time t. For any admissible control $\mathbf{u}$, let $\mathbb{P}^{\mathbf{u}}_{\mathbf{x}_t,t}$ be the corresponding controlled path law. By Girsanov's theorem (under the usual linear growth/Lipschitz assumptions ensuring absolute continuity), we have

$$D_{\text{KL}}\Big(\mathbb{P}^{\mathbf{u}}_{\mathbf{x}_t,t}\,\Big\|\,\mathbb{P}^{\text{ref}}_{\mathbf{x}_t,t}\Big) = \mathbb{E}_{\mathbb{P}^{\mathbf{u}}_{\mathbf{x}_t,t}}\left[\int_t^1 \tfrac{1}{2}\,\|\mathbf{u}(\mathbf{x}_s, s)\|^2\, ds\right].$$

Hence the KL–regularized continuous-time RL objective (Equation (7)) can be written as

$$V^*(\mathbf{x}_t, t, \eta) = \sup_{\mathbf{u}}\Big\{\eta\, \mathbb{E}_{\mathbb{P}^{\mathbf{u}}_{\mathbf{x}_t,t}}\big[r(\mathbf{x}_1)\big] - D_{\text{KL}}\big(\mathbb{P}^{\mathbf{u}}_{\mathbf{x}_t,t}\,\|\,\mathbb{P}^{\text{ref}}_{\mathbf{x}_t,t}\big)\Big\}. \tag{20}$$

According to Equation (17), we have

$$V^*(\mathbf{x}_t, t, \eta) = \mathbb{E}_{\mathbf{x}_1\sim p_{\text{ref}}(\mathbf{x}_1|\mathbf{x}_t,t)}\big[\exp\big(\eta\, r(\mathbf{x}_1)\big)\big]\,,$$

which yields the claimed closed form. The expression depends only on the reward distribution given $\mathbf{x}_t$, so the coefficient $1/\eta$ (KL temperature) can be tuned *post hoc* without retraining. □

E.2. Proof of Lemma 2

Proof. Consider the reverse-time SDE on $[t_k, t_k{+}\Delta]$:

$$d\mathbf{x}_t = \Big(f(\mathbf{x}_t, t) + \sigma(t)^2\, s_\theta(\mathbf{x}_t, t)\Big)\, dt + \sigma(t)\, d\mathbf{B}_t, \qquad \mathbf{x}_{t_k} \text{ given.}$$

Assume the usual Lipschitz and linear-growth conditions so that the Euler–Maruyama (EM) scheme is valid. Over a small step $\Delta = \frac{L}{T}$, the EM update at $(\mathbf{x}_{t_k}, t_k)$ reads

$$\mathbf{x}_{t_k+\Delta} \;\approx\; \mathbf{x}_{t_k} + \Big(f(\mathbf{x}_{t_k}, t_k) + \sigma(t_k)^2\, s_\theta(\mathbf{x}_{t_k}, t_k)\Big)\Delta + \sigma(t_k)\big(\mathbf{B}_{t_k+\Delta} - \mathbf{B}_{t_k}\big).$$

Since $\mathbf{B}_{t_k+\Delta} - \mathbf{B}_{t_k} \sim \mathcal{N}(\mathbf{0},\, \Delta\,\mathbf{I})$ and is independent of $\mathbf{x}_{t_k}$, conditioning on the same start $(\mathbf{x}_{t_k}, t_k)$ implies

$$\mathbf{x}_{t_k+\Delta}\,\big|\,\mathbf{x}_{t_k} \;\approx\; \mathcal{N}\Big(\underbrace{\mathbf{x}_{t_k} + \big(f(\mathbf{x}_{t_k}, t_k) + \sigma(t_k)^2 s_\theta(\mathbf{x}_{t_k}, t_k)\big)\Delta}_{\mu_{t_k}}\,,\; \underbrace{\sigma(t_k)^2\,\Delta\,\mathbf{I}}_{\Sigma_{t_k}}\Big).$$

Standard EM error bounds give $\|\mathbb{E}[\varphi(\mathbf{x}_{t_k+\Delta})] - \mathbb{E}[\varphi(\mathbf{x}^{\text{EM}}_{t_k+\Delta})]\| = \mathcal{O}(\Delta)$ for smooth test functions φ (weak order 1), and $\mathbb{E}\|\mathbf{x}_{t_k+\Delta} - \mathbf{x}^{\text{EM}}_{t_k+\Delta}\|^2 = \mathcal{O}(\Delta)$ (strong order $1/2$), so the Gaussian approximation holds to first order in Δ. □

E.3. Proof of Theorem 3

We decompose the proof into three parts: the characterization of the segment-level optimum, the inductive propagation of optimal marginals, and the base case at $t = 0$.

Define the *desirability* (exponentiated value) at any $(\mathbf{x}, t)$ by

$$z^*(\mathbf{x}, t; \eta) \;:=\; \exp\big(V^*(\mathbf{x}, t, \eta)\big) \;=\; \mathbb{E}_{\mathbf{x}_1 \sim p_{\mathrm{ref}}(\cdot|\mathbf{x},t)}\big[\exp\big(\eta\, r(\mathbf{x}_1)\big)\big]\,.$$

By the Markov property of p_{ref} and the law of total expectation, z^* satisfies the *multiplicative Bellman equation* for any intermediate time t and next time $t{+}\Delta$:

$$z^*(\mathbf{x}, t; \eta) \;=\; \mathbb{E}_{\mathbf{x}' \sim p_{\mathrm{ref}}(\cdot|\mathbf{x},t)}\big[z^*(\mathbf{x}', t{+}\Delta; \eta)\big], \qquad V^*(\mathbf{x}, t, \eta) \;=\; \log \mathbb{E}_{\mathbf{x}' \sim p_{\mathrm{ref}}(\cdot|\mathbf{x},t)}\big[e^{V^*(\mathbf{x}', t+\Delta, \eta)}\big]. \tag{21}$$

Lemma 5 (Segment-level VRPO optimum)**.** *Fix a state $(\mathbf{x}, t)$ and a short segment to $t{+}\Delta$. The VRPO optimum with temperature β satisfies*

$$p_\theta(\mathbf{x}' \mid \mathbf{x}, t) \;=\; p_{\mathrm{ref}}(\mathbf{x}' \mid \mathbf{x}, t)\, \exp\Big(\tfrac{1}{\beta} V^*(\mathbf{x}', t{+}\Delta, 1/\beta) - \tfrac{1}{\beta} V^*(\mathbf{x}, t, 1/\beta)\Big). \tag{22}$$

That is, the optimal conditional distribution is the Doob h-transform of the reference dynamics tilted by the terminal reward.

Proof. The VRPO loss with temperature β is

$$\mathcal{L}_{\mathrm{VRPO}} = -\,\mathbb{E}\left[\log \mathrm{sigmoid}\left(\beta \log \frac{p_\theta(\mathbf{x}'^{+} \mid \mathbf{x}, t)}{p_{\mathrm{ref}}(\mathbf{x}'^{+} \mid \mathbf{x}, t)} - \beta \log \frac{p_\theta(\mathbf{x}'^{-} \mid \mathbf{x}, t)}{p_{\mathrm{ref}}(\mathbf{x}'^{-} \mid \mathbf{x}, t)}\right)\right],$$

with preferences $\mathbf{x}'^{+} \succ \mathbf{x}'^{-}$ iff $V^*(\mathbf{x}'^{+}, t{+}\Delta, 1/\beta) > V^*(\mathbf{x}'^{-}, t{+}\Delta, 1/\beta)$. Bradley–Terry analysis implies the optimal ratio has the form

$$\log \frac{p_\theta(\mathbf{x}' \mid \mathbf{x}, t)}{p_{\mathrm{ref}}(\mathbf{x}' \mid \mathbf{x}, t)} = C(\mathbf{x}, t) + \tfrac{1}{\beta} V^*(\mathbf{x}', t{+}\Delta, 1/\beta).$$

Normalizing over $\mathbf{x}'$ and using Equation (21), gives (22). □

Lemma 6 (Inductive propagation of marginals)**.** *Define the tilted marginal*

$$p^*(\mathbf{x}, t) \;\propto\; p_{\mathrm{ref}}(\mathbf{x}, t)\, \exp\big(\tfrac{1}{\beta} V^*(\mathbf{x}, t, 1/\beta)\big). \tag{23}$$

If $p_\theta(\mathbf{x}, t) = p^*(\mathbf{x}, t)$ at time t, then applying the conditional kernel (22) yields $p_\theta(\mathbf{x}', t{+}\Delta) = p^*(\mathbf{x}', t{+}\Delta)$.

Proof. Starting from $p_\theta(\mathbf{x}, t) = p^*(\mathbf{x}, t)$, propagate one segment:

$$\begin{aligned}
p_\theta(\mathbf{x}', t{+}\Delta) &= \int p_\theta(\mathbf{x}' \mid \mathbf{x}, t)\, p_\theta(\mathbf{x}, t)\, d\mathbf{x} \\
&\propto \int p_{\mathrm{ref}}(\mathbf{x}' \mid \mathbf{x}, t)\, \exp\Big(\tfrac{1}{\beta} V^*(\mathbf{x}', t{+}\Delta, 1/\beta) - \tfrac{1}{\beta} V^*(\mathbf{x}, t, 1/\beta)\Big) \\
&\qquad \cdot\, p_{\mathrm{ref}}(\mathbf{x}, t)\, \exp\big(\tfrac{1}{\beta} V^*(\mathbf{x}, t, 1/\beta)\big)\, d\mathbf{x} \\
&= \exp\big(\tfrac{1}{\beta} V^*(\mathbf{x}', t{+}\Delta, 1/\beta)\big)\, p_{\mathrm{ref}}(\mathbf{x}', t{+}\Delta) \;\propto\; p^*(\mathbf{x}', t{+}\Delta).
\end{aligned}$$

Thus the tilted form is preserved. □

Lemma 7 (Base case under independence (Domingo-Enrich et al., 2025))**.** *Suppose that under the reference process the initial and terminal states are independent, i.e., $p_{\mathrm{ref}}(\mathbf{x}_1 \mid \mathbf{x}_0, 0) = p_{\mathrm{ref}}(\mathbf{x}_1, 1)$. Then the optimal value satisfies*

$$V^*(\mathbf{x}_0, 0, \eta) = \log \mathbb{E}_{\mathbf{x}_1 \sim p_{\mathrm{ref}}(\cdot,1)}\big[e^{\eta r(\mathbf{x}_1)}\big] =: C_\eta, \qquad V^*(\mathbf{x}_1, 1, \eta) = \eta\, r(\mathbf{x}_1).$$

In particular, $V^*(\mathbf{x}_0, 0, \eta)$ is a constant (independent of $\mathbf{x}_0$).

Proof. By definition of the value (closed form),

$$V^*(\mathbf{x}_0, 0, \eta) = \log \mathbb{E}_{\mathbf{x}_1 \sim p_{\mathrm{ref}}(\cdot|\mathbf{x}_0, 0)}\big[e^{\eta r(\mathbf{x}_1)}\big].$$

Under the independence assumption, $p_{\mathrm{ref}}(\mathbf{x}_1 \mid \mathbf{x}_0, 0) = p_{\mathrm{ref}}(\mathbf{x}_1, 1)$, hence

$$V^*(\mathbf{x}_0, 0, \eta) = \log \mathbb{E}_{\mathbf{x}_1 \sim p_{\mathrm{ref}}(\cdot, 1)}\big[e^{\eta r(\mathbf{x}_1)}\big] =: C_\eta,$$

which does not depend on $\mathbf{x}_0$. At the terminal time, the boundary condition gives $V^*(\mathbf{x}_1, 1, \eta) = \eta\, r(\mathbf{x}_1)$. □

Lemma 8 (Base case and terminal marginal)**.** *At $t = 0$, the initial distribution of the learned process coincides with that of the reference process:*

$$p_\theta(\mathbf{x}, 0) \;=\; p_{\mathrm{ref}}(\mathbf{x}, 0). \tag{24}$$

Proof. Since $\mathbf{x}_0$ and $\mathbf{x}_1$ are independent under the reference process, Lemma 7 implies

$$p_\theta(\mathbf{x}, 0) \;\propto\; p_{\mathrm{ref}}(\mathbf{x}, 0)\, \exp(C_\eta),$$

where C_η is a constant independent of $\mathbf{x}$. Normalization then yields (24). □

Theorem 3 (Optimality of VRPO)**.** *Consider a trajectory on $\mathcal{T}_T$ that is decomposed into K short segments, each of length L with $\Delta = \frac{L}{T}$ sufficiently small. If VRPO is applied to segment-level preferences defined by $V_k^*(\mathbf{x}_t, t, \eta)$, then under sufficient model capacity, the resulting model is equivalent to training directly on the optimal target distribution*

$$p^*(\mathbf{x}_1) \;\propto\; p_{data}(\mathbf{x}_1)\, \exp\big(\eta\, r(\mathbf{x}_1)\big). \tag{15}$$

Proof of Theorem 3. Combining Lemmas 5, 6, and 8, we conclude that stage-wise VRPO training yields a terminal distribution

$$p(\mathbf{x}_1) \;\propto\; p_{\mathrm{data}}(\mathbf{x}_1)\, \exp(\eta r(\mathbf{x}_1)),$$

which is exactly the optimal target. Moreover, it is straightforward to verify that the intermediate distributions also coincide. □

Table 4. Results of quantitative comparison with baselines using the preference comparison prompts.

Methods	PickScore	imageReward	Aesthetic score	HPSv2
SDXL	22.2645	0.6738	6.4385	0.2786
SPO	23.3715	1.0234	6.8181	0.2909
Ours	**23.6443**	**1.1823**	**7.2582**	**0.2916**

F. Detailed Comparison of Training Time

As shown in Table 5, compared to the standard DPO methods (taking DiffusionDPO as an example), VRPO adds an extra stage, but the time spent is significantly less than theirs. The purpose of the additional first stage is to train a step-wise value function, which allows VRPO to perform DPO at intermediate timesteps during diffusion image generation, rather than just after the image is fully generated. This approach saves a significant amount of time during the second stage of diffusion training.

Furthermore, we discovered that training just a prediction head of the step-wise value function yields very good results. We save the embeddings from before the prediction head, so the time required for this stage is much less than SPO.

Table 5. Comparison of training time between baseline methods.

Methods/Stage	Value Function Training	Diffusion Training
Diffusion-DPO-SD1.5	-	384 A100 hours
SPO-SD1.5	92 A100 hours	44 A100 hours
VRPO-SD1.5	0.3 A6000 hours	62 A100 hours
Diffusion-DPO-SDXL	-	4800 A100 hours
SPO-SDXL	152 A100 hours	124 A100 hours
VRPO-SDXL	0.5 A6000 hours	174 A100 hours

G. More Ablation Studies

Segment Length. We conduct experiments on different segment lengths, and the results are shown in Table 6. As the segment length increases from 1 to 4, performance consistently improves across nearly all metrics. This confirms that aggregating preference signals over multiple timesteps is highly beneficial. Beyond length 4, further increases yield diminishing returns and eventual degradation.

While our theoretical results motivate using sufficiently short segments so that the endpoint preference is a good approximation of the trajectory preference, the empirical optimum at length 4 reflects a crucial practical trade-off between *approximation fidelity* and *preference informativeness*. When the segment length is too small (e.g., 1 step), the two candidate endpoints within a segment are often very similar, leading to a small reward gap between the winning and losing samples. This makes the induced pairwise preference signal for the DPO objective weaker and more susceptible to noise. As the segment length increases moderately, the reward difference becomes more distinguishable, yielding more informative preference pairs and a stronger training signal. However, if the segment becomes too long, the short-segment approximation becomes less accurate. Length 4 achieves the best balance between these two factors and is therefore used in all our main experiments.

To empirically support this explanation, we measured the absolute reward gap between the winning and losing samples on the Pick-a-Pic test split using our reward model. As shown in Table 7, shorter segments exhibit significantly smaller mean reward gaps and fewer pairs with distinct reward differences. The informativeness increases and balances well with approximation fidelity around length 4, which perfectly aligns with the downstream performance trend observed in Table 6.

Table 6. Results of ablation studies on segment length.

Segment Length	PickScore↑	ImageReward↑	Aesthetic↑	HPSv2↑
1	22.8005	1.1912	6.7166	0.2805
2	22.8557	1.1925	7.0675	0.2828
3	22.8909	**1.2698**	<u>7.0767</u>	0.2829
4	**23.0656**	<u>1.2404</u>	**7.1894**	<u>0.2867</u>
6	<u>23.0293</u>	1.2208	7.0426	**0.2875**
8	22.9653	1.1933	6.8451	0.2850

Table 7. Absolute reward gap between winning and losing samples across different segment lengths on the Pick-a-Pic test split.

Segment Length	Mean Gap	Gap > 0.01 (%)	Gap > 0.015 (%)
1	0.0140	46.50	33.18
2	0.0160	50.25	37.00
3	0.0182	51.93	39.21
4	0.0198	56.60	45.50
6	0.0197	56.75	44.12
8	0.0214	57.00	44.00

Miss-ranking and Approximation Error. Although precisely controlling the degree of miss-ranking is challenging, we

provide a detailed ablation study by evaluating multiple value models with varying degrees of accuracy. As shown in Table 8, we trained multiple classification heads (50 classes) under different random seeds and training budgets, and then evaluated them on a test split of 1,000 unseen trajectories. We report two losses from Algorithm 2 (CRPS and cross-entropy) as well as the Top-1 and Top-3 accuracy of the predicted rank bucket to measure the miss-ranking degree.

Beyond the miss-ranking study, we directly evaluate the approximation error of the learned value $V_\phi(x_t, t, \eta)$ against a Monte Carlo estimate of the true value:

$$V^*(x_t, t, \eta) = \log \mathbb{E}_{x_1 \sim p_{\text{ref}}(\cdot|x_t,t)} \big[\exp(\eta r(x_1)) \big], \tag{25}$$

which corresponds to Eq (11) in the main paper. Concretely, on the Pick-a-Pic test prompts, we use SDXL as the reference model and perform $K = 8$ Monte Carlo rollouts per state (x_t, t). We report the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the absolute approximation error in Table 8.

Sensitivity to Value-Estimation Noise. Based on the results in Table 8, we observe a consistent overall trend: improved value-model accuracy (i.e., lower MAE/RMSE/CRPS and higher classification accuracy) leads to better downstream VRPO performance, with the strongest value model achieving the best final generative results across all reward metrics (PickScore, ImageReward, Aesthetic, and HPS). At the same time, the downstream generation performance degrades gradually, rather than catastrophically, as the value-model accuracy decreases. This indicates that our proposed method exhibits practical robustness to moderate value-estimation noise.

Table 8. Results of ablation studies on value estimation accuracy (miss-ranking and approximation error) and its impact on downstream generation performance.

MAE↓	RMSE↓	CRPS↓	CE↓	Top-1 Acc↑	Top-3 Acc↑	PickScore↑	ImgRew↑	Aesthetic↑	HPS↑
0.0535	0.0642	0.02475	2.007	26.40%	55.00%	22.8583	1.1398	6.8556	0.2833
0.0503	0.0629	*0.02070*	*1.910*	*30.40%*	*63.00%*	*22.9602*	1.1729	6.9367	*0.2841*
0.0497	**0.0602**	0.02293	1.999	28.80%	58.60%	22.8824	**1.2816**	*7.0424*	0.2829
0.0487	*0.0605*	**0.01631**	**1.692**	**36.50%**	**72.50%**	**23.0656**	*1.2404*	**7.1894**	**0.2867**

Bias Introduced by Discretized CDF Modeling. Another potential source of value-estimation error is the bias introduced by discretizing the continuous value distribution into discrete classification bins. This issue is partially reflected in Sec. 5.3 and Table 1 of the main paper. In particular, a coarse discretization (e.g., 20 or 30 bins) introduces noticeable bias, which causes a clear drop in downstream generation performance. However, we observe that the performance becomes much more stable once the number of bins reaches approximately 50 and above. These empirical findings demonstrate that while discretization bias is non-negligible at low resolutions, it is substantially and effectively reduced at high resolutions without compromising the overall framework.

H. Detailed Prompts

Prompts for Figure 1 are provided in Table 9. Prompts for Figure 3 are provided in Table 10.

I. MLLM Evaluation Prompts

The prompts used for evaluation with ChatGPT-5 were as follows:

> Please evaluate and compare two images based on the following prompt and three dimensions:
>
> Prompt associated with the images: {prompt}
>
> Evaluation dimensions: 1. Overall performance: Comprehensive assessment of image quality, expressiveness and how well it fulfills the prompt 2. Visual appeal: Evaluation of the aesthetics, color matching and composition of the image 3. Text image alignment: Assessment of how well the image content matches the provided prompt
>
> For each dimension, please clearly judge:
>
> - Image 1 is better
> - Image 2 is better

Table 9. Prompts for Figure 1.

Row & Line	Prompt
Row 1, Line 1	A beautiful natural woman.
Row 1, Line 3	A drawing of a funny giraffe eating from a tree.
Row 1, Line 4	Red and white high tops shoes and Bitcoin.
Row 1, Line 5	A golden retriever representing god.
Row 1, Line 6	Throne, dark scene, moonlight.
Row 1, Line 7	Highway to hell that goes into a mountain covered in snow.
Row 2, Line 3	Cute girl, Kyoto animation, 4k, high resolution.
Row 2, Line 4	Corgi with helmet on bicycle.
Row 2, Line 5	Fire sorcerer.
Row 2, Line 6	A huskey playing footbal on the beach.
Row 3, Line 1	Tripod legged robot in a lake.
Row 3, Line 2	A robot that is made out of wood.
Row 3, Line 3	Raccoon with a sombrero riding on a atv with Tequila in hand, photorealism.
Row 3, Line 4	Hello kitty mecha, gears of war, style Artstation, octane render, unreal engine 6, epic game Graphics, Fantasy, cyberpunk, conceptual art, Ray tracing.
Row 3, Line 5	A gloomy rabbit drinks wine.

Table 10. Prompts for Figure 3.

Line	Prompt
Line 1	A beautiful Indian woman by the beach.
Line 2	A man eating a glazed donut and a woman eating a chocolate cake.
Line 3	Beautiful dancing unicorns.
Line 4	A full-body shot of a beautiful woman wearing a dress, with sharp focus on her eyes, depicted in an elegant, intricate style by artgerm, jason chan, and mark hill.
Line 5	Cowby angel standing in a field of flowers.
Line 6	A cake with butterflies and rainbows.
Line 7	A coloring book page of a lotus flower, white clear background.
Line 8	A bee devouring the world.

- Both are equal

Please answer in a clear and concise format, with each dimension on a separate line, for example:

- Overall performance: Image 1 is better
- Visual appeal: Both are equal
- Text image alignment: Image 2 is better

Although we allowed ChatGPT-5 to assign an "equal" score in the overall performance dimension, it never selected this option during evaluation.

J. The Roadmap of VRPO Methodology

Our method is organized around a two-stage construction that converts an image-domain terminal reward into a segment-level training signal for diffusion preference optimization. In the first stage, we learn a value-distribution function that estimates the conditional distribution of the terminal reward obtained by completing the reverse diffusion process from an intermediate state (x_t, t) under the reference model. This distributional object is important because the terminal reward $r(x_1)$ is only observed at the final image, while VRPO requires a reliable return signal for intermediate trajectory segments. Given the learned value distribution, we then induce the optimal value function by taking the log moment-generating function with respect to the reward coefficient η. This converts the learned conditional reward distribution into a scalar optimal value $V^*(x_t, t, \eta)$, which serves as a principled return for comparing short diffusion segments. In the second stage, we decompose a long diffusion trajectory into short segments and apply VRPO at the segment level: candidate endpoints are ranked

Table 11. Compact notation summary.

Notation	Description
x_t, $x_{0:1}$	Diffusion state at time t and the full trajectory.
x_1	Terminal generated sample.
p_{ref}, p_θ	Reference process and trainable diffusion model.
$\pi_\theta(x_1)$	Terminal marginal induced by p_θ.
$s_\theta(x,t)$	Trainable score function.
$\nabla_x \log p_{\mathrm{ref}}(x,t)$	Reference score function.
$u(x,t)$	Control variable, i.e., scaled score gap.
$r(x_1)$	Terminal image-domain reward.
η, β	Reward coefficient and DPO temperature, with $\beta = 1/\eta$.
$Z(x_t,t)$	Random reward $r(x_1)$ under $x_1 \sim p_{\mathrm{ref}}(x_1 \mid x_t,t)$.
$F_\phi(z \mid x_t,t)$	Learned CDF of the conditional reward distribution.
$q_\phi(z \mid x_t,t)$	Value-distribution induced by F_ϕ.
$V^*(x_t,t,\eta)$	Optimal value function from SOC.
$V^*_\phi(x_t,t,\eta)$	Learned optimal value computed from F_ϕ.
$z^*(x,t;\eta)$	Desirability function $\exp(V^*(x,t,\eta))$.
T, $\mathcal{T}_T$	Number of diffusion steps and discrete time grid.
L, Δ	Segment length and segment duration $\Delta = L/T$.
K, t_k	Number of segments and start time of the k-th segment.
x^+, x^-	Preferred and dispreferred segment endpoints.
$\mathcal{L}_{\mathrm{VRPO}}$	Segment-level VRPO loss.

according to the induced optimal value, and the diffusion model is updated using a DPO-style likelihood-ratio objective on these short transitions. This design avoids treating a single trajectory likelihood as the terminal marginal probability, while preserving consistency with the global image-domain reward. Table 11 is a compact notation table summarizing the main distributions, value functions, and time-scale variables, to reduce symbol-switching of VRPO.

K. Human User Study

In addition to the automatic preference evaluation with MLLM, we conduct a user study to further assess whether the preference improvements are consistent with human judgments. We randomly sample 20 image groups from the evaluation set used for the AI-based assessment. Each group contains images generated by SDXL, SPO, and VRPO for the same prompt. For each group, the three images are presented in a randomized order, and human raters are asked to select the best image according to three criteria: visual quality, text–image alignment, and overall performance. The final results are reported as win rates over all human selections.

As shown in Table 12, VRPO achieves the highest win rate across all three criteria. Specifically, VRPO obtains a win rate of 56.5% on visual quality, 52.5% on text–image alignment, and 51.5% on overall performance, outperforming both SDXL and SPO by a clear margin. These results suggest that the improvements brought by our value-as-return framework are not limited to automatic evaluators, but are also reflected in human preferences.

Table 12. Human user study among SDXL, SPO, and VRPO on 20 randomly sampled image groups. Results are reported as win rates (%) over all human selections.

Criterion	SDXL	SPO	VRPO
Visual Quality	13.0	30.5	**56.5**
Text Alignment	16.5	31.0	**52.5**
Overall Performance	14.0	34.5	**51.5**