arXiv:2402.08265v2 [cs.CV] 12 May 2024

A Dense Reward View on Aligning Text-to-Image Diffusion with Preference

Shentao Yang [*1] **Tianqi Chen** [*1] **Mingyuan Zhou** [1]

Abstract

Aligning text-to-image diffusion model (T2I) with preference has been gaining increasing research attention. While prior works exist on directly optimizing T2I by preference data, these methods are developed under the bandit assumption of a latent reward on the entire diffusion reverse chain, while ignoring the sequential nature of the generation process. This may harm the efficacy and efficiency of preference alignment. In this paper, we take on a finer dense reward perspective and derive a tractable alignment objective that emphasizes the initial steps of the T2I reverse chain. In particular, we introduce temporal discounting into DPO-style explicit-reward-free objectives, to break the temporal symmetry therein and suit the T2I generation hierarchy. In experiments on single and multiple prompt generation, our method is competitive with strong relevant baselines, both quantitatively and qualitatively. Further investigations are conducted to illustrate the insight of our approach. Source code is available at https://github.com/Shentao-YANG/Dense_Reward_T2I.

1. Introduction

Text-to-image diffusion model (T2I, Ramesh et al., 2022; Saharia et al., 2022), trained by large-scale text-image pairs, has achieved remarkable success in image generation. As an effort towards more helpful and less harmful generations, methods have been proposing to align T2I with preference, partially motivated by the progress of human/AI-feedback alignment for large language models (LLMs) (Bai et al., 2022b; OpenAI, 2023; Touvron et al., 2023). Prior works in this field typically optimize the T2I against an explicit reward function trained in the first place (Wu et al., 2023b; Xu et al., 2023; Lee et al., 2023b). To remove the complexity in the modeling and computing of an explicit reward function, recent work has generalized direct preference optimization (DPO, Rafailov et al., 2023) from LLM into T2I's preference alignment (Wallace et al., 2023a), under a counterpart assumption of DPO that there is a latent reward function evaluating the entire diffusion reverse chain as a whole.

While DPO-style approaches have shown impressive potential, from the reinforcement learning (RL) perspective, these methods typically formulate the diffusion reverse chain as a contextual bandit, *i.e.*, treating the entire generation trajectory as a single action; though the diffusion reverse chain is intrinsically a sequential generation process (Sohl-Dickstein et al., 2015; Ho et al., 2020). Since the reverse chain typically requires tens or even thousands of steps (Song et al., 2020; 2021), such a bandit assumption, in particular, of a reward function on the whole chain/trajectory, can lead to a combinatorially large decision space over all timesteps. This issue is twined with the well-known sparse reward (delayed feedback) issue in RL (Andrychowicz et al., 2017; Liu et al., 2019), where an informative feedback is only provided after generating the entire trajectory. We hereafter use "sparse reward" to refer to this issue. Without considering the sequential nature of the generation process, it is known from RL and LLM literature that this sparse reward issue, which often comes with high gradient variance and low sample efficiency (Guo et al., 2022), can clearly hurt model training (Marbach & Tsitsiklis, 2003; Takanobu et al., 2019).

In this paper, we contribute to the research on DPO-style explicit-reward-free alignment methods by taking on a finer-grain *dense-reward* perspective, motivated by recent studies on the latent preference-generating reward function in NLP (*e.g.*, Yang et al., 2023) and robotics (*e.g.*, Kim et al., 2023; Hejna et al., 2023). Instead of the hypothetical *trajectory-level* reward function, we assume a latent reward function that can score *each step* of the reverse chain, in hoping an easier learning problem from the RL viewpoint (*e.g.*, Laidlaw et al., 2023). Inspired by studies on diffusion and T2I generation that the initial portion of the reverse chain sets up the image outline based on the given text conditional, and image's high-level attributes and aesthetic shapes (Ho et al., 2020; Wang & Vastola, 2023), we hypothesize that emphasizing those initial steps in T2I's preference alignment can help efficacy and efficiency, since those steps can be

[*]Equal contribution [1]The University of Texas at Austin. Correspondence to: Shentao Yang <shentao.yang@mccombs.utexas.edu>, Mingyuan Zhou <mingyuan.zhou@mccombs.utexas.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

more directly related to the final de-noised image's being the preferred one. Under this hypothesis, we break the temporal symmetry in the DPO-style alignment losses by introducing the temporal discounting factor, a key RL ingredient, into T2I's alignment. Practically, we develop a lower bound of the resulting Bradley-Terry preference model (Bradley & Terry, 1952), which leads to a tractable loss to train a T2I for preference alignment in an explicit-reward-free manner.

We test our method on the task of single prompt generation, which is easier for investigation; and the more challenging multiple prompt generation, where we align our T2I with the preference pertaining to one set of prompts and evaluate on another large-scale set of prompts. On both tasks, our method exhibits competitive quantitative and qualitative performance against strong baselines. We conduct further studies on the effectiveness of emphasizing the initial steps of the reverse chain in T2I's alignment, which to our best knowledge has not been well investigated in literature.

2. Main Method

2.1. Notations and Assumptions

In this section, we state the notations and assumptions for deriving our method. As discussed in Section 1, our first and foremost assumption is a latent *dense* reward.

Assumption 2.1. There is a latent reward function $r(s_t, a_t)$ that can score each step t of the T2I reverse chain.

We adopt the notations in prior works (*e.g.*, Fan et al., 2023; Black et al., 2023) to formulate the diffusion reverse process under the conditional generation setting as an Markov decision process (MDP), specified by $\mathcal{M} = (\mathbb{S}, \mathbb{A}, \mathcal{P}, r, \gamma, \rho)$. Specifically, let π_θ be the T2I with trainable parameters θ, *i.e.*, the policy network; $\{\boldsymbol{x}_t\}_{t=T}^0$ be the diffusion reverse chain of length T; and $\boldsymbol{c}$ be the text conditional, *i.e.*, the conditioning variable. We have, $\forall\, t$,

$$\begin{aligned} s_t &\triangleq (\boldsymbol{x}_t, t, \boldsymbol{c})\,, & \pi_\theta(a_t \,|\, s_t) &\triangleq p_\theta(\boldsymbol{x}_{t-1} \,|\, \boldsymbol{x}_t, t, \boldsymbol{c}), \\ a_t &\triangleq \boldsymbol{x}_{t-1}\,, & \rho(s_0) &\triangleq (\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \delta(T), \delta(\boldsymbol{c})), \\ \mathcal{P}(s_{t+1} \,|\, s_t, a_t) &\triangleq \delta(\boldsymbol{x}_{t-1}, t-1, \boldsymbol{c})\,, & r(s_t, a_t)\,,\ \gamma &\in [0, 1], \end{aligned}$$

where $\delta(\cdot)$ is the delta measure and $\mathcal{P}(\cdot \,|\, s_t, a_t)$ is a deterministic transition. We denote generically the reverse chain generated by a T2I under the text conditional $\boldsymbol{c}$ as a trajectory τ, *i.e.*, $\tau \triangleq (s_0, a_0, s_1, a_1, \ldots, s_T) \iff (\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_0) \,|\, \boldsymbol{c}$. Note that for notation simplicity, $\boldsymbol{c}$ is absorbed into the state part of τ.

Similar to Wallace et al. (2023a), we consider the setting where we are given two trajectories (reverse chains) with equal length T. For simplicity, assume that τ^1 is the better one, *i.e.*, $\tau^1 \succ \tau^2$. Let tuple $\mathrm{ord} \triangleq (1, 2)$ and $\sigma(\cdot)$ denotes the sigmoid function, *i.e.*, $\sigma(x) = 1/(1 + \exp(-x))$.

As in standard RL settings (Sutton & Barto, 2018; Yang et al., 2022b), the reward function r in $\mathcal{M}$ needs to be bounded. Without loss of generality, we assume $r(s, a) \in [0, 1]$, and thus $r(s, a)$ may be interpreted as the probability of satisfying the preference when taking action a at state s.

Assumption 2.2. In $\mathcal{M}$, $\forall\, (s, a) \in \mathbb{S} \times \mathbb{A}$, $0 \le r(s, a) \le 1$.

The performance of a (generic) policy π is typically evaluated by the expected cumulative discounted rewards (Sutton & Barto, 2018), which is defined as,

$$\eta(\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^T \gamma^t\, r(s_t, a_t) \,|\, s_0 \sim \rho, a_t \sim \pi, s_{t+1} \sim \mathcal{P}\right]. \quad (1)$$

Assumption 2.3. Based on Eq. (1), we assume that for a (generation) trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$, its quality is evaluated by $e(\tau) \triangleq \sum_{t=0}^T \gamma^t\, r(s_t, a_t)$.

Remark 2.4 (Practical Rationality of $e(\tau)$). Since the final step of the reverse chain depends on all previous steps, a score or (human) evaluation on the final de-noised image should indeed evaluate the whole corresponding de-noising chain, *i.e.*, the entire generation trajectory τ. This notion is particularly intuitive when the de-noising process is deterministic, *e.g.*, DDIM (Song et al., 2020). Though motivated from the RL viewpoint (Eq. (1)), in evaluating a T2I's generation, typically humans first check its *conceptual* shapes and matching with the text prompt; and if that's OK, then look at finer details in the image. Thus, the initial steps of the reverse chain, which set up image *outlines* (Section 1), can play a more important role in an image's being preferred. This insight is distilled into $e(\tau)$ by using $\gamma < 1$, which emphasizes the contribution from the initial steps.

2.2. Method Derivation

The derivation of our method is inspired by the RL literature (*e.g.*, Kakade & Langford, 2002; Schulman et al., 2015; Peng et al., 2019) and DPO (Rafailov et al., 2023). Due to the space limit, in this section we only present the key steps. A step-by-step derivation is deferred to Appendix B.

Directly optimizing $\eta(\pi)$ in Eq. (1) requires constantly sampling from the current learning policy, which can be less practical for T2I's preference alignment. We are therefore motivated by the cited literature to consider an approximate off-policy objective. Specifically, we employ the initial pre-trained T2I, denoted as π_I; and generate the off-policy trajectories by some "old" policy π_O, where π_O may be chosen as π_I or some saved policy checkpoint not far from π_I. We denote $d_{\pi_O}(s)$ as the stationary distribution of π_O (detailed in Appendix B.2.1). To avoid generating unnatural images, we impose a KL regularization towards π_I on the learning policy π. Together, we arrive at the following regularized policy optimization problem

$$\begin{aligned} \arg\max_\pi \quad & \mathbb{E}_{s\sim d_{\pi_O}(s)} \mathbb{E}_{a\sim\pi(a\,|\,s)}\left[r(s, a)\right] \\ & - C \cdot \mathbb{E}_{s\sim d_{\pi_O}(s)}\left[D_{\mathrm{KL}}\left(\pi(\cdot\,|\,s) \,\|\, \pi_I(\cdot\,|\,s)\right)\right] \quad (2) \\ \text{s.t.} \quad & \int_{\mathbb{A}} \pi(a\,|\,s)\,\mathrm{d}a = 1, \quad \forall\, s \in \mathbb{S}\,, \end{aligned}$$

where C is a tuning regularization/KL coefficient.

By solving the first-order condition of the Lagrange form of Eq. (2), we can get the optimal (regularized) policy π^* as

$$\pi^*(a \mid s) = \exp\left(r(s,a)/C\right) \pi_I(a \mid s) \,/ Z(s) \ , \tag{3}$$

where $Z(s)$ denotes the partition function, taking the form

$$Z(s) = \int_{\mathbb{A}} \exp\left(r(s,a)/C\right) \pi_I(a \mid s) \, \mathrm{d}a \, .$$

We also have the relation between π^* and r, as

$$r(s,a) = C \log\left[\pi^*(a \mid s)/\pi_I(a \mid s)\right] + C \log Z(s) \, . \tag{4}$$

For a *given* trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$, after plugging in Eq. (4), $e(\tau)$ can be expressed by π^* as

$$e(\tau) = C \sum_{t=0}^{T} \left[\gamma^t \log \frac{\pi^*(a_t \mid s_t)}{\pi_I(a_t \mid s_t)}\right] + C \log Z(\tau) \, . \tag{5}$$

where we denote $\log Z(\tau) \triangleq \sum_{t=0}^{T} \gamma^t \log Z(s_t)$ for notation simplicity since the discounted sum is over all $s_t \in \tau$.

Under the Bradley-Terry (BT) model, by plugging in Eq. (5), the probability of ord under $\{e(\tau^k)\}_{k=1}^2$ and hence π^* is

$$\begin{aligned} &\Pr\left(\mathrm{ord} \,|\, \pi^*, \{e\left(\tau^k\right)\}_{k=1}^2\right) = \sigma\left(e\left(\tau^1\right) - e\left(\tau^2\right)\right) \\ &= \frac{\exp\left(C \sum_{t=0}^{T} \gamma^t \log \frac{\pi^*\left(a_t^1 \mid s_t^1\right)}{\pi_I\left(a_t^1 \mid s_t^1\right)}\right) Z\left(\tau^1\right)^C}{\sum_{i=1}^{2} \exp\left(C \sum_{t=0}^{T} \gamma^t \log \frac{\pi^*\left(a_t^i \mid s_t^i\right)}{\pi_I\left(a_t^i \mid s_t^i\right)}\right) Z\left(\tau^i\right)^C} \, . \end{aligned} \tag{6}$$

Eq. (6), however, contains the intractable partition functions $Z(\tau^1)$ and $Z(\tau^2)$. We will provide a tractable lower bound of Eq. (6) by arguing that $Z(\tau^1) \geq Z(\tau^2)$. Our argument is based on the reward-shaping technique (Ng et al., 1999).

Definition 2.5 (Reward Shaping)**.** A shaping-reward function Φ is a real-valued function on the state space, $\Phi : \mathbb{S} \to \mathbb{R}$. It induces a new MDP $\mathcal{M}' = (\mathbb{S}, \mathbb{A}, \mathcal{P}, r', \gamma, \rho)$ where $r'(s,a) \triangleq r(s,a) + \Phi(s)$.

Lemma 2.6 (Invariance of Optimal Policy under Reward Shaping)**.** *The optimal (regularized) policy Eq. (3) under the reward-shaped MDP $\mathcal{M}'$ is the same as that in the original MDP $\mathcal{M}$.*

The proof is deferred to Eq. (19) in Appendix B.2.2. Note that $\mathcal{M}'$ and $\mathcal{M}$ share the same state and action space. Thus, it makes sense to consider the invariance of the optimal policy, where invariance means at each state taking the same action with the same probability.

Definition 2.7. The equivalence class $[r]$ of the reward function r is the set of all reward functions that can be obtained from r by reward shaping, *i.e.*, $\forall\, r' \in [r], \exists\, \Phi : \mathbb{S} \to \mathbb{R}, \ s.t.\ r'(s,a) - r(s,a) = \Phi(s), \forall\, s \in \mathbb{S}, a \in \mathbb{A}$.

Remark 2.8. By Lemma 2.6, all reward functions in $[r]$ share the same optimal (regularized) policy as r, *i.e.*, Eq. (3).

We are now able to justify our argument: $Z(\tau^1) \geq Z(\tau^2)$.

Theorem 2.9. *Under Assumption 2.2, and a sufficiently large regularization coefficient C, for any finite number $K \geq 2$ of trajectories $\{\tau^k\}_{k=1}^K$ where $\tau^1 \succ \tau^2 \succ \cdots \succ \tau^K$, $\exists\, r' \in [r]$, $s.t.$, $Z(\tau^1) \geq Z(\tau^2) \geq \cdots \geq Z(\tau^K)$ under r'.*

We defer the proof of Theorem 2.9 to Appendix B.3.2.

Remark 2.10. For the value of C, as we will see in the proof, we technically require that $\forall (s,a) \in \mathbb{S} \times \mathbb{A}, r(s,a)/C \leq \mathrm{const} \approx 1.79$. Under Assumption 2.2, $C \geq 0.56$ will suffice. We note that this technical requirement helps reducing the search space of the hyperparameter C in practice.

Remark 2.11. By Lemma 2.6, r' in Theorem 2.9 and the original r lead to the same optimal policy π^*, which is our ultimate target. Due to this invariance, for notation simplicity, we hereafter refer to r' as r, though we *may* actually work in the "equivalent" MDP $\mathcal{M}' = (\mathbb{S}, \mathbb{A}, \mathcal{P}, r', \gamma, \rho)$.

With Theorem 2.9, we can provide a simpler lower bound to $\Pr(\mathrm{ord} \,|\, \pi^*, \{e\left(\tau^k\right)\}_{k=1}^2)$ in Eq. (6),

$$\Pr\left(\mathrm{ord} \,|\, \pi^*, \{e\left(\tau^k\right)\}_{k=1}^2\right) \geq \frac{\exp\left(C \sum_{t=0}^{T} \gamma^t \log \frac{\pi^*\left(a_t^1 \mid s_t^1\right)}{\pi_I\left(a_t^1 \mid s_t^1\right)}\right)}{\sum_{i=1}^{2} \exp\left(C \sum_{t=0}^{T} \gamma^t \log \frac{\pi^*\left(a_t^i \mid s_t^i\right)}{\pi_I\left(a_t^i \mid s_t^i\right)}\right)} \, . \tag{7}$$

Recall that $e(\tau)$ evaluates a trajectory τ's quality, and thus a better trajectory comes with a higher $e(\tau)$. Hence $\Pr(\mathrm{ord} \,|\, \pi^*, \{e(\tau^k)\}_{k=1}^2) = \max \Pr(\cdot \,|\, \pi^*, \{e(\tau^k)\}_{k=1}^2)$, *i.e.*, under $\mathcal{M}$ with π_I and conditioning on π^*, ord should be the most probable ordering under the BT model shown in Eq. (6). Thus, in order to approximate π^*, we train π_θ by maximizing the lower bound Eq. (7) of the *corresponding* BT likelihood of ord, which leads to the negative-log-likelihood *loss* function for training π_θ as

$$\begin{aligned} &\mathcal{L}_\gamma(\theta \,|\, \mathrm{ord}, \{e(\tau^k)\}_{k=1}^2) \\ &= -\log \sigma\left(C \mathbb{E}_{t \sim \mathrm{Cat}(\{\gamma^t\})}\left[\log \frac{\pi_\theta\left(a_t^1 \mid s_t^1\right)}{\pi_I\left(a_t^1 \mid s_t^1\right)} - \log \frac{\pi_\theta\left(a_t^2 \mid s_t^2\right)}{\pi_I\left(a_t^2 \mid s_t^2\right)}\right]\right) \, , \end{aligned} \tag{8}$$

where $\mathrm{Cat}(\{\gamma^t\})$ denotes the categorical distribution on $\{0, \ldots, T\}$ with the probability vector $\{\gamma^t / \sum_{t'} \gamma^{t'}\}_{t=0}^T$ and C is overloaded to absorb the normalization constant.

Interpretation. To see what $\mathcal{L}_\gamma(\theta \,|\, \mathrm{ord}, \{e(\tau^k)\}_{k=1}^2)$, our loss in Eq. (8), is doing, let's calculate its gradient.

Since Eq. (8) is an objective for minimization problem, the gradient update direction is $-\nabla_\theta \mathcal{L}_\gamma$. For notation simplicity, we denote $\widetilde{e}(\tau^k) \triangleq C \sum_{t=0}^{T} \gamma^t \log \frac{\pi_\theta(a_t^k \mid s_t^k)}{\pi_I(a_t^k \mid s_t^k)}$. We have

$$\begin{aligned} \frac{\partial\left(-\mathcal{L}_\gamma(\theta \,|\, \mathrm{ord}, \{e(\tau^k)\}_{k=1}^2)\right)}{\partial \theta} &= \underbrace{\frac{\exp\left(\widetilde{e}\left(\tau^2\right) - \widetilde{e}\left(\tau^1\right)\right)}{1 + \exp\left(\widetilde{e}\left(\tau^2\right) - \widetilde{e}\left(\tau^1\right)\right)}}_{(*)} \times C \\ &\times \sum_{t=0}^{T} \gamma^t \left(\pi_I(a_t^1 \mid s_t^1) \nabla_\theta \log \pi_\theta(a_t^1 \mid s_t^1) - \pi_I(a_t^2 \mid s_t^2) \nabla_\theta \log \pi_\theta(a_t^2 \mid s_t^2)\right) . \end{aligned} \tag{9}$$

Algorithm 1 Outline of Our Off-policy Learning Routine.

Input: Prompt distribution $p(\boldsymbol{c})$, T2I π_θ, training steps M_{tr}, trajectory collect period M_{col}, # prompts to collect trajectories N_{pr}, # trajectories for each prompt N_{traj}.
Initialization: Sample N_{pr} prompts $\{\boldsymbol{c}\} \sim p(\boldsymbol{c})$, get N_{traj} trajectories for each $\boldsymbol{c}$.
for iter $\in \{1, \dots, M_{\text{tr}}\}$ **do**
  Sample a mini-batch $\mathcal{B} \triangleq \{(\tau_i^1, \tau_i^2)_{\boldsymbol{c}_i}\}_i$ from storage.
  Optimize π_θ via Eq. (8) using $\mathcal{B}$.
  if iter $\%$ $M_{\text{col}} == 0$ **then**
    Re-sample N_{pr} prompts $\{\boldsymbol{c}\} \sim p(\boldsymbol{c})$, get N_{traj} trajectories for each $\boldsymbol{c}$, and update the storage.
  end if
end for

Detailed derivation is in Appendix B.3.1. The term (*) is high when $\widetilde{e}(\tau^2) > \widetilde{e}(\tau^1)$, *i.e.*, in the unwanted case where the discounted (relative) likelihood of the inferior trajectory τ^2 is higher. In that case, we increase the likelihood of $(s_t, a_t) \in \tau^1$ and decrease $(s_t, a_t) \in \tau^2$. Note that this mechanism is weighted by γ^t, with which we emphasize the earlier steps in the reverse chain. As discussed in Section 1, this could be more effective in getting desirable final images.

Additionally, for (s_t, a_t), if $\pi_I(a_t \mid s_t)$ is small, our changes (increase or decrease likelihood) can be small too. This may be interpreted as those (s_t, a_t) are at the edge of the initial distribution threatening the generation of realistic images. Meanwhile, if $\pi_I(a_t \mid s_t)$ is high, our changes can be also high, since we now have more "room" for improving and our gradient utilizes this to achieve safe and effective training.

2.3. Practical Implementation

In practice, we assume that π_θ is optimized over a given prompt distribution $p(\boldsymbol{c})$, where $p(\boldsymbol{c}) = \delta(\boldsymbol{c})$ if we fine-tune π_θ on a single prompt $\boldsymbol{c}$, and $p(\boldsymbol{c}) = \text{Unif}(\mathcal{D}(\boldsymbol{c}))$ for a dataset $\mathcal{D}(\boldsymbol{c})$ of prompts if tuning π_θ on multiple prompts.

We implement our algorithm as an online off-policy learning routine. Similar to prior RL works (*e.g.*, Mnih et al., 2013; Lillicrap et al., 2016), we iterate between **(1)** using the current π_θ to sample N_{traj} trajectories for each of the N_{pr} prompts sampled from $p(\boldsymbol{c})$; and **(2)** training π_θ via Eq. (8) on mini-batches of trajectories sampled from all stored. To mimic the classical RLHF settings (*e.g.*, Ziegler et al., 2019; Ouyang et al., 2022), we set $N_{\text{traj}} = 5 \geq 2$ for resource efficiency. In calculating the loss Eq. (8), we sample N_{step} timesteps from $\text{Cat}(\{\gamma^t\})$ to estimate the expectation inside $\sigma(\cdot)$. Algo. 1 outlines the key steps of our method.

2.4. Connection with the DPO Objective.

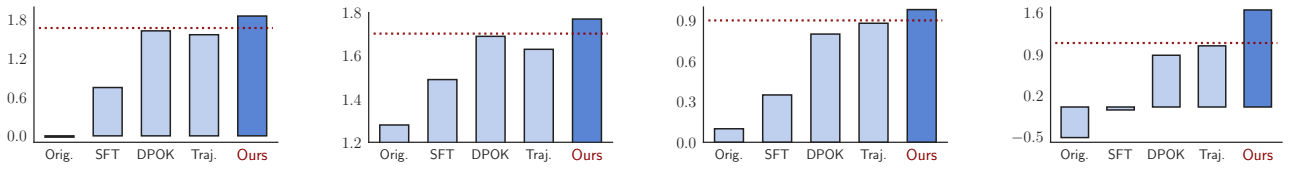
The original DPO loss (Eq. (7) in Rafailov et al. (2023)) can be obtained as a variant of Eq. (8) when setting $\gamma = 1$, after factorizing out the probability at each step t. Using $\gamma = 1$ in our formulation is equivalent to the DPO-style trajectory-level bandit setting since $\gamma = 1$ makes the contribution of each timestep t to $e(\tau)$ *symmetric*, *i.e.*, each timestep is equally important, and therefore each timestep t is *symmetric* in the loss as well. Likewise, in DPO-style trajectory-level bandit setting, since the trajectory as a *whole* receives a single reward, this reward/evaluation does not distinguish each step t within the trajectory either, making each timestep t *symmetric* again in the training loss, same as our variant with $\gamma = 1$. Due to this connection in the loss, we refer to this variant as "trajectory-level reward," indistinguishable to whether it actually comes from a trajectory-level bandit setting or our formulation but with $\gamma = 1$. As a reminder, in our formulation, if we set $\gamma < 1$, then the contribution of each timestep t to $e(\tau)$ will not be symmetric, since earlier steps will be emphasized. This leads to the desirable *asymmetry* of timestep t in the loss, as shown in Eq. (8).

3. Related Work

T2I's Alignment with Preference. There have been growing interests in aligning T2I's, or more broadly diffusion models', generations to (human) preferences. Efforts have been putting on tuning the models on curated data (Podell et al., 2023; Dai et al., 2023) or re-captioning existing image datasets (Betker et al., 2023; Segalis et al., 2023), to bias T2I generation towards better text fidelity and aesthetics. These data enhancement efforts may complement our method.

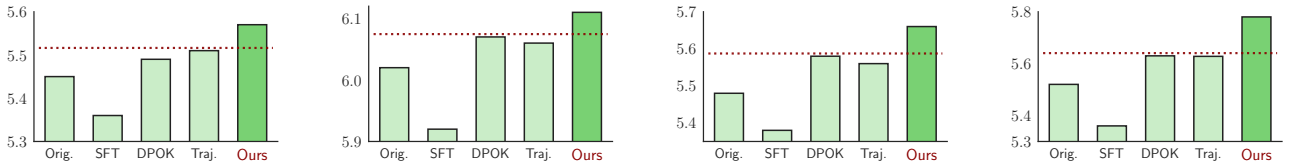
To more directly optimize the feedback, methods have been proposing to fine-tune T2I with respect to (*w.r.t.*) reward models pre-trained on large-scale human preference datasets (Xu et al., 2023; Wu et al., 2023a; Kirstain et al., 2023). Lee et al. (2023a) and Wu et al. (2023b) adapt the classical supervised training by fine-tuning T2I via reward-weighted likelihood or discarding low-reward images, with online versions extended by Dong et al. (2023). By formulating the denoising process as an MDP, policy gradient methods are adopted to fine-tune T2I for specific rewards (Fan & Lee, 2023; Fan et al., 2023; Black et al., 2023) or polishing the input prompts (Hao et al., 2022). Further assuming a *differentiable* reward function, a more direct alignment/feedback-optimization can be achieved by backpropagating the reward function's gradient through the reverse chain (*e.g.*, Clark et al., 2023; Prabhudesai et al., 2023; Wallace et al., 2023b). Although optimizing *w.r.t.* explicit rewards have shown efficacy and efficiency, it requires a stronger assumption than our method on having an explicit scalar reward function, while assuming analytic gradients of the reward function is even stronger. By contrast, our method only requires binary comparison between generated images/trajectories, which is among the simplest in T2I's preference alignment.

Most close to our work, Diffusion-DPO (Wallace et al., 2023a) also considers an explicit-reward-free T2I alignment



(a) Color (b) Count (c) Composition (d) Location

Figure 1: ImageReward scores for the seen prompts in the single prompt experiments. "Orig." denotes the original SD1.5. "SFT" is the supervised fine-tuned model. "Traj." denotes the classical DPO-style objective discussed in Section 2.4, *i.e.*, assuming trajectory-level reward. All our produced results are the average over 100 samples. Horizontal line indicates the best baseline result.



(a) Color (b) Count (c) Composition (d) Location

Figure 2: Aesthetic scores for the seen prompts in the single prompt experiments. Number reporting and abbreviations follow Fig.1.

method. It is nevertheless developed under a different setting where the generation latents are discarded, and thus it needs to approximate the reverse process with the forward. However, given the relatively-small scale of the preference alignment stage, storing the reverse chains can be both feasible and straightforward. We thereby eschew such an approximation and use the exact generation latents. More importantly, as in DPO (Rafailov et al., 2023), Diffusion-DPO is derived by assuming reward on the *whole* chain/trajectory, obtainable as a variant of our method (Section 2.4) and distinct from our dense reward perspective. In experiments, we validate the efficacy of our perspective by comparing with this approach of "trajectory-level reward."

Appendix E reviews literature on **(1)** dense *v.s.* sparse training guidance for sequential generative models, **(2)** characterizing the (latent) preference generation distribution, and **(3)** learning-from-preference in related fields.

4. Experiments

To straightforwardly evaluate our method's ability to satisfy preference, motivated by recent papers in directly tuning T2I *w.r.t.* pre-trained rewards (Section 3) and relevant papers in NLP (*e.g.*, Ramachandran et al., 2021; Feng et al., 2023; Yang et al., 2023), in our experiments, we use the following logics: We obtain preference among multiple trajectories by some open-source scorer trained on data of human preference over T2I's generations; and test our method's ability in increasing the score, as an indication of the model's improved alignment with (human) preference. The scorer factors in text fidelity. For preference simulation, given a prompt and N_{traj} corresponding images, the higher the score, the more preferable the image is.

For computational efficiency, our policy π_θ is implemented as LoRA (Hu et al., 2021) added on the U-net (Ronneberger et al., 2015) module of a frozen pre-trained Stable Diffusion v1.5 (SD1.5, Rombach et al., 2022), and we only train the LoRA parameters. With SD1.5, the generated images are of resolution 512×512. For all our main results, we set the discount factor γ to be $\gamma = 0.9$. We perform ablation study on the γ value in Section 4.3 **(b)**. As in prior works (*e.g.*, Fan et al., 2023; Black et al., 2023), in both sampling trajectories and generating evaluation images, we use DDPM sampler with 50 inference steps and classifier-free guidance (Ho & Salimans, 2022). We use the default guidance scale of 7.5. Source code is publicly released.

Table 1: Seen and unseen prompts in Section 4.1 for each domain.

Domain	Seen	Unseen
Color	A green colored rabbit.	A green colored cat.
Count	Four wolves in the park.	Four birds in the park.
Composition	A cat and a dog.	A cat and a cup.
Location	A dog on the moon.	A lion on the moon.

4.1. Single Prompt

Settings. To facilitate investigation, we first test our method on the single-text-prompt setting in DPOK (Fan et al., 2023), *i.e.*, using one prompt during LoRA fine-tuning. As in DPOK, the goal is to test our method on training the policy T2I to achieve generating objects with specified colors, counts, or locations, or generating composition of two objects. We borrow the seen (training) and unseen prompts from DPOK, which are tabulated in Table 1. In all single prompt experiments, we use the explicit reward model in DPOK, ImageReward (Xu et al., 2023), to generate preference. We report both ImageReward and (Laion) Aesthetic score (Schuhmann et al., 2022), averaged over 100 generated images.

Implementation. For a fair comparison, we collect the same total amount of 20000 images/trajectories as DPOK. Rather than its fully-online image-collection strategy, we are motivated by recent RLHF works (*e.g.*, Ziegler et al., 2019;

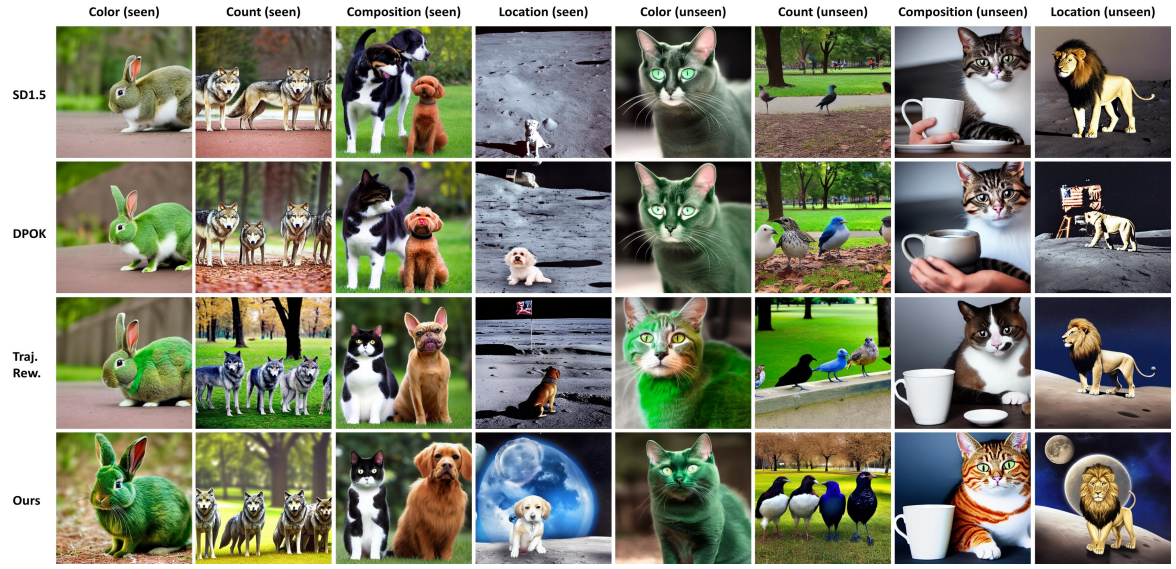


Figure 3: Generated images in the single prompt experiment for both seen and unseen prompts (Table 1). Each comparison is generated from the same random seed. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Stiennon et al., 2020; Bai et al., 2022a) to more practically divide our trajectory collection into four stages, where each stage collects 5000 trajectories and discards the previously collected ones. As in DPOK, we use LoRA with rank 4 and train the model for a total of $M_{\text{tr}} = 10000$ steps, and hence $M_{\text{col}} = 2500$ steps, $N_{\text{pr}} = 1000$. We set the KL coefficient $C = 10$ and $N_{\text{step}} = 3$. Section 4.3 **(c)** ablates the value of C. More details on hyperparameters are in Appendix F.1.

Results. We compare our method with the the original SD1.5 ("Orig."), supervised fine-tuned model ("SFT"), DPOK, and the classical DPO-style objective, *i.e.*, the approach of assuming trajectory-level reward, which is abbreviated as "Traj.". As discussed in Section 2.4, "Traj." can be obtained by setting $\gamma = 1$ in our loss Eq. (8). We follow the DPOK paper to plot the ImageReward in Fig. 1 and Aesthetic score in Fig. 2 for the seen prompts, where the results for "Orig.", "SFT", and DPOK are directly from the DPOK paper. Fig. 3 shows examples of the generated images from both our method and the baselines. More image comparisons are deferred to Appendix G.1.

As shown in Fig. 1 and Fig. 2, our method can improve both ImageReward, the preference generating metric, and the unseen Aesthetic score. The higher scores of our method over DPOK on both metrics validate the efficacy of our method for T2I's preference alignment. Comparing with "Traj.", our method improves more over the original SD1.5, which we attribute to our dense reward perspective, implemented by introducing temporal discounting to emphasize the initial steps of the diffusion reverse chain. From Fig. 3, it is clear that, on both seen and unseen text prompts, our method generates images that are not only faithfully matched with the prompts, but also of higher aesthetic quality, *e.g.*, having more colorful details and/or backgrounds. Section 4.3 **(a)** compares the generation trajectories of our method and the baselines. Indeed, our method generates the desired shapes earlier, which explains why it produces better final images.

4.2. Multiple Prompts

Settings. We consider a more challenging setting where we apply our method to train a T2I on the HPSv2 (Wu et al., 2023a) train prompts and evaluate on the HPSv2 test prompts, which have no intersection with the train prompts. We obtain preference by HPSv2 and report the average of both HPSv2 and Aesthetic score over all HPSv2 test prompts. Due to the large test-set size (3200 prompts), we follow the HPSv2 paper to generate one image per prompt for evaluation.

Implementation. We use the same trajectory-collection strategy as in the single prompt experiments (Section 4.1). Due to the task complexity and the large size of the HPSv2 train set ($> 100,000$ prompts), we collect a total of $100,000$ trajectories, divided into ten collection stages. Each stage collects $10,000$ trajectories and discards the previously collected ones. We use LoRA with rank 32 and train the model for a total of $M_{\text{tr}} = 40,000$ steps, and hence $M_{\text{col}} = 4000$ steps, $N_{\text{pr}} = 2000$. We set the KL coefficient $C = 12.5$ and ablates the value of C in Section 4.3 **(c)**. We use $N_{\text{step}} = 1$ based on compute constraints such as GPU memory. Appendix F.2 provides more hyperparameter settings.

Results. Table 2 shows the HPSv2 and Aesthetic score for our method and selected relevant and/or strong baselines from the HPSv2 paper, with the full set of baselines deferred to Table 4 of Appendix A. All baselines available in HPSv2 Github Repository are directly cited. As in Section 4.1,

Table 2: HPSv2 and Aesthetic score for the multiple prompt experiment. Shown here are results for selected relevant and/or strong baselines, with full set of results in Table 4 of Appendix A. The first four result columns are the four styles in HPSv2 test set and "Average" is the overall average. Best result in each metric is bold. Note that HPSv2 paper and Github repository do not report Aesthetics score.

Model	Animation	Concept-art	Painting	Photo	Average	Aesthetic
DALL·E 2	27.34	26.54	26.68	27.24	26.95	-
Stable Diffusion v1.5	27.43	26.71	26.73	27.62	27.12	5.62
Stable Diffusion v2.0	27.48	26.89	26.86	27.46	27.17	-
SDXL Refiner 0.9	28.45	27.66	27.67	27.46	27.80	-
Dreamlike Photoreal 2.0	28.24	27.60	27.59	27.99	27.86	-
Trajectory-level Reward	29.37	28.81	28.83	29.16	29.04	5.94
Ours	**30.46**	**29.95**	**30.01**	**29.93**	**30.09**	**6.31**

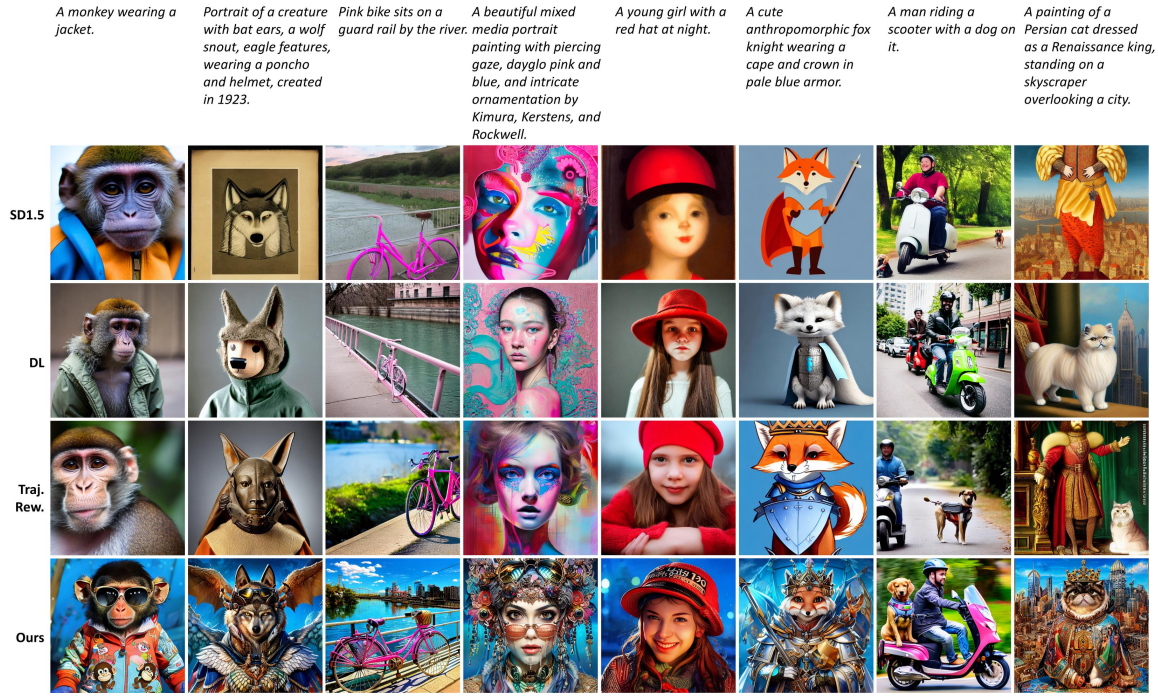


Figure 4: Generated images in the multiple prompt experiment from our method and baselines, with prompts. "DL" denotes Dreamlike Photoreal 2.0, the best baseline from HPSv2 paper. "Traj. Rew." is the classical DPO-style objective of assuming trajectory-level reward.

we further compare with the classical DPO-style objective of assuming trajectory-level reward (Section 2.4). Fig. 4 shows examples of generated images from our method and baselines, with more image comparisons in Appendix G.2.

As seen in Table 2, our method is able to improve the preference generating metric, HPSv2, and the unseen Aesthetic score. The improvement from our method is larger than the variant of assuming trajectory-level reward, validating our insight of emphasizing the initial part of the T2I generation process, a product of our distinct dense reward perspective. In Fig. 4, we see that our method generates images well matched with the text prompts, in some cases better than the baselines, *e.g.*, on the prompts of "a girl at night," "fox knight," and "scooter with a dog on." From both short and the more challenging long prompts, our method is able to generate vivid images, often with sophisticated aesthetic shapes. Together with the image examples in Appendix G.2, Fig. 4 qualitatively validates the efficacy of our method.

4.3. Further Study

This section considers the following four research questions to better understand our method.

(a): *Does the T2I trained by our method indeed generate the desired shapes earlier in the diffusion reverse chain?*

As discussed in Section 1, we hypothesize that emphasizing the initial steps of the T2I generation trajectory can help the effectiveness and efficiency of preference alignment. As a verification, Fig. 5 digs into the generated images of the prompt "A green colored rabbit." in the single prompt experiment, by showing the generation trajectories corre-

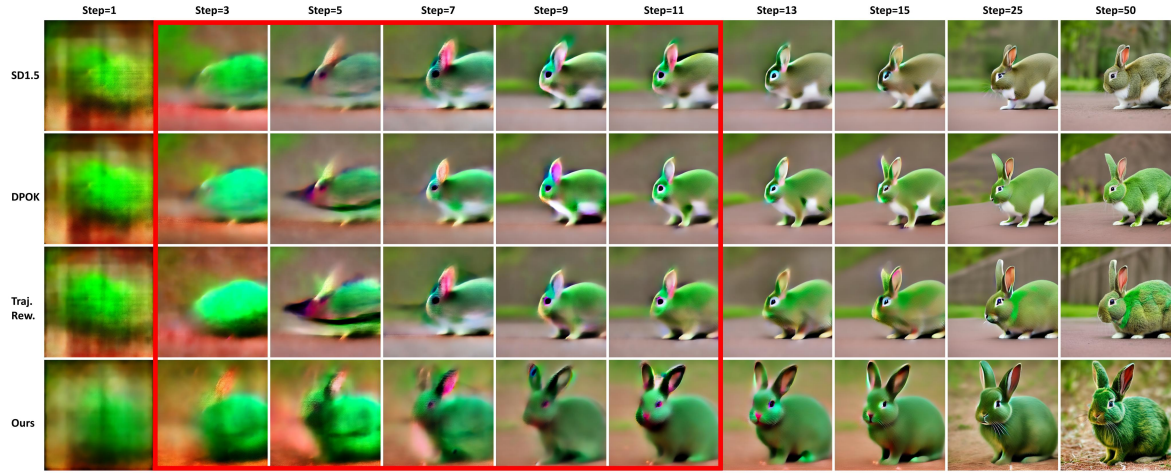
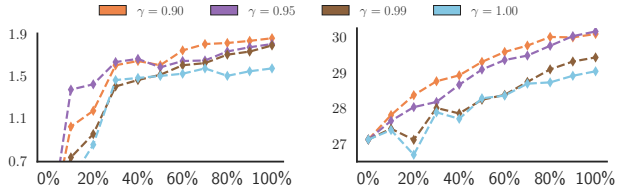
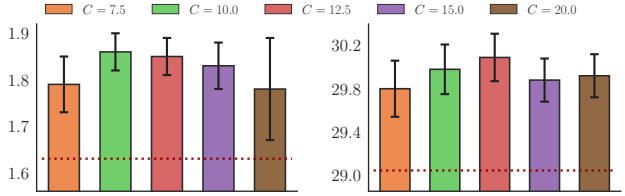


Figure 5: Generation trajectories from our method and the baselines on the prompt "A green colored rabbit." in the single prompt experiment, correspond to the images in Fig. 3. Shown are the $\hat{\boldsymbol{x}}_0$ predicted from the latents at the specified steps of the reverse chain.



(a) Single (ImageReward) (b) Multiple (HPSv2)

Figure 6: Preference generating metrics over the training process, for the single and multiple prompt experiments under various discount factor γ. x-axis represents $t\%$ of the training process. In **(a)** all lines start from -0.02 at 0%, the value of "Orig." in Fig. 1a.



(a) Single (ImageReward) (b) Multiple (HPSv2)

Figure 7: Preference generating metrics with error bars showing one standard deviation, for the single and multiple prompt experiments under various KL coefficient C. Horizontal line indicates the best baseline result from Fig. 1a and Table 2.

sponding to the images in Fig. 3. Specifically, we compare our method and the baselines on the images $\hat{\boldsymbol{x}}_0$ predicted from the latents at the specified timesteps of the reverse chain. More trajectory comparisions are in Appendix G.3.

As shown in Fig. 5, and in particular the steps circled out by the red rectangle therein, our method can generate identifiable shapes of a rabbit as early as at Steps 3 and 5, while the baselines are still largely unrecognizable, *e.g.*, similar to a mouse. At step 11, our method is able to produce a relatively complete image to the given prompt, while the baselines are much cruder. This comparison confirms that, with the incorporation of $\gamma < 1$, our method can match the given prompt earlier in the reverse chain, and thereby more steps later in the chain can be allocated to polish pictorial details and aesthetics, leading to better/preferable final images.

(b): *What will happen if we change the value of γ?*

To investigate the impact of temporal discount factor γ on training T2I for preference alignment, we consider more values of γ between $\gamma = 0.9$ used in our main results, and $\gamma = 1$ in the classical approach of trajectory-level reward. Fig. 6 plots the preference generating metrics over the training process, under $\gamma \in \{0.9, 0.95, 0.99, 1.0\}$, for the single prompt ("A green colored rabbit.") and multiple prompt experiments. We use the same evaluation protocols as in the main results. For HPSv2, we plot the average over the test set. Patterns on other single prompts are similar.

As shown in Fig. 6, using a smaller temporal discount factor, such as $\gamma = 0.9$ or $\gamma = 0.95$, trains T2I faster and better, compared to larger γ values, especially the classical DPO-style loss of $\gamma = 1$. Recall from Section 1 that a smaller γ emphasizes more on the initial part of the reverse chain, while a sparse trajectory-level reward, equivalent to $\gamma = 1$, can incur training instability. In Fig. 6, on both experiments, $\gamma = 0.9$ or $\gamma = 0.95$ generally leads to larger improvement at the beginning of the training process. This validates our intuition and prior study that stressing the earlier steps of the reverse chain could improve the training efficiency of aligning T2I with preference. From Fig. 6, even using $\gamma = 0.99$, a small break on the temporal symmetry in the DPO-style losses, can improve training efficiency and stability over the classical setting of $\gamma = 1$. This further corroborates the efficacy of our dense reward perspective on T2I's alignment. Appendix C further discusses the effect of γ on training T2I.

(c): *Is our method robust to the choice of KL coefficient C?*

To study the sensitivity of our method to the KL coefficient C in our loss Eq. (8), we vary the value of C from the values set in Sections 4.1 and 4.2. Fig. 7 plots the scores of the preference generating metrics for experiments in the single

Table 3: Human evaluation on the multiple prompt experiment. Shown are our "win rate" against the baselines specified in Fig. 4, *i.e.*, the percentage of times our method is preferred in binary comparisons. Detailed description on the setup is in Appendix F.3.

Opponent	SD1.5	Dreamlike	Traj. Rew.
Win Rate	76.8%	68.3%	65.1%

prompt ("A green colored rabbit.") and multiple prompts. Other single prompts show similar patterns. For HPSv2, we again plot the average over the test set, with Aesthetic and breakdown scores for each style in Table 5 at Appendix A.

From Fig. 7, we see that our method is generally robust across a range of KL coefficient C. A small value of C may be prone to overfitting while a large value may distract/slow the training process, both of which deteriorate the results.

(d): *Are the images from our method preferred by humans?*

To further verify our method, we collect human evaluations on the generated images in the multiple prompt experiment, where binary comparisons between two images from two models are conducted. Table 3 shows the "win rate" of our method over each of the baselines in Fig. 4. Detailed setups of the human evaluation are provided in Appendix F.3.

The preference for our method over each baseline is evident in Table 3. Recall that the preference source, HPSv2 scorer, is trained on human preference data. The gain of our method over raw SD1.5 verifies the efficacy of our method in aligning T2I with preference. Further, images from our method are more often preferred over the corresponding images from the classical trajectory-level reward approach. This again validates our dense reward perspective that introduces temporal discounting into T2I's preference alignment.

5. Conclusion

To suit the explicit-reward-free preference-alignment loss to the sequential generation nature of T2I and improve on the classical trajectory-level reward assumption, in this paper, we take on a dense reward perspective and introduce temporal discounting into the alignment objective, motivated by both an easier learning task in RL and the generation hierarchy of T2I reverse chain. By experiments and further studies, we validate the efficacy of our method and reveal its key insight. Future work may involve extending our method to noisy preference labels and applying it to broader applications, such as text-to-video or image-to-image generation.

Impact Statement

Our paper contributes to the ongoing research on increasing helpfulness and decreasing harmfulness of generative models, by proposing a method that seeks to improve the efficacy and efficiency of aligning T2I with preference. Of a special note, our method does not require training an explicit reward model, which can potentially save some compute and resources. On the other hand, as prior preference alignment methods, it is possible that our method will be misused to train malicious T2I by aligning with some unethical or ill-intended preference. This potential negative impact may be alleviated by a more closer monitoring on the datasets and preference sources to which our method is applied.

Limitations

As with classical off-policy RL and RLHF methods, our method's iteration between model training and data collection incurs additional complexity and costs, compared to the pure offline approach of gathering data only once prior to policy training. On the other hand, it is known that off-policy methods can reduce the mismatch between learning policy's generation distribution and the data distribution, and generally lead to more stable training and better results than pure offline methods. Another limitation of our method is that our method requires storing the generation reverse chains. Though this is feasible and straightforward given the relatively-small scale of the preference alignment stage, our approach does raise extra CPU-memory and/or storage requirements, compared to only storing the final images and discarding all generation latents. As an example, in our experiments with SD1.5, storing the generation latents requires about two times more CPU memory (*not* GPU memory), calculated as $50 \times 4 \times 64^2/(512^2 \times 3) \times (16/8) \approx 2.08$, where the last multiplier comes from the fact that our generation latents are stored in `bfloat16` format and the final images are in `uint8`. This limitation may be further alleviated by using a more advanced diffusion/T2I sampler.

Acknowledgements

The authors acknowledge the support of NSF-IIS 2212418, NIH-R37 CA271186, McCombs REG, and TACC. S. Yang acknowledges the support of the University of Texas Graduate Continuing Fellowship.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1ANxQW0b.

Akrour, R., Schoenauer, M., and Sebag, M. Preference-based policy learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, pp. 12–27. Springer, 2011.

Akrour, R., Schoenauer, M., and Sebag, M. April: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pp. 116–131. Springer, 2012.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2:3, 2023.

Bıyık, E., Lazar, D. A., Sadigh, D., and Pedarsani, R. The green choice: Learning and influencing human decisions on shared roads. In *2019 IEEE 58th conference on decision and control (CDC)*, pp. 347–354. IEEE, 2019.

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

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Brown, D., Goo, W., Nagarajan, P., and Niekum, S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Brown, D. S., Goo, W., and Niekum, S. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, pp. 330–359. PMLR, 2020.

Castricato, L., Havrilla, A., Matiana, S., Pieler, M., Ye, A., Yang, I., Frazier, S., and Riedl, M. Robust preference learning for storytelling via contrastive reinforcement learning. *arXiv preprint arXiv:2210.07792*, 2022.

Chen, H., Liu, X., Yin, D., and Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Clark, K., Vicol, P., Swersky, K., and Fleet, D. J. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Dai, X., Hou, J., Ma, C.-Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E. P., and Hu, Z. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv*, abs/2304.06767, 2023. URL https://api.semanticscholar.org/CorpusID:258170300.

Ethayarajh, K., Xu, W., Jurafsky, D., and Kiela, D. Human-centered loss functions (halos). Technical report, Contextual AI, 2023. URL https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf.

Fan, Y. and Lee, K. Optimizing ddpm sampling with shortcut fine-tuning. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256415971.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023.

Feng, Y., Yang, S., Zhang, S., Zhang, J., Xiong, C., Zhou, M., and Wang, H. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. In *The Eleventh International Conference on Learning Representations*, 2023.

Finn, C., Christiano, P. F., Abbeel, P., and Levine, S. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *ArXiv*, abs/1611.03852, 2016.

Fürnkranz, J., Hüllermeier, E., Cheng, W., and Park, S.-H. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89:123–156, 2012.

Guo, H., Tan, B., Liu, Z., Xing, E., and Hu, Z. Efficient (soft) q-learning for text generation with limited good data. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6969–6991, 2022.

Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., and Wang, J. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Hao, Y., Chi, Z., Dong, L., and Wei, F. Optimizing prompts for text-to-image generation. *ArXiv*, abs/2212.09611, 2022. URL https://api.semanticscholar.org/CorpusID:254853701.

Hejna, D. J. and Sadigh, D. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pp. 2014–2025. PMLR, 2023a.

Hejna, J. and Sadigh, D. Inverse preference learning: Preference-based rl without a reward function. *arXiv preprint arXiv:2305.15363*, 2023b.

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., and Sadigh, D. Contrastive prefence learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*, 2023.

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.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, À., Jones, N. J., Gu, S., and Picard, R. W. Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog. *ArXiv*, abs/1907.00456, 2019.

Jaques, N., Shen, J. H., Ghandeharioun, A., Ferguson, C., Lapedriza, A., Jones, N., Gu, S. S., and Picard, R. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*, 2020.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kim, C., Park, J., Shin, J., Lee, H., Abbeel, P., and Lee, K. Preference transformer: Modeling human preferences using transformers for RL. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Peot1SFDX0.

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.

Knox, W. B., Hatgis-Kessell, S., Booth, S., Niekum, S., Stone, P., and Allievi, A. Models of human preference for learning reward functions. *ArXiv*, abs/2206.02231, 2022. URL https://api.semanticscholar.org/CorpusID:249395243.

Knox, W. B., Hatgis-Kessell, S., Adalgeirsson, S. O., Booth, S., Dragan, A. D., Stone, P., and Niekum, S. Learning optimal advantage from preferences and mistaking it for reward. *ArXiv*, abs/2310.02456, 2023. URL https://api.semanticscholar.org/CorpusID:263620440.

Korbak, T., Shi, K., Chen, A., Bhalerao, R., Buckley, C. L., Phang, J., Bowman, S. R., and Perez, E. Pretraining language models with human preferences. *arXiv preprint arXiv:2302.08582*, 2023.

Kwan, W.-C., Wang, H., Wang, H., and Wong, K.-F. A survey on recent advances and challenges in reinforcement learningmethods for task-oriented dialogue policy learning. *arXiv preprint arXiv:2202.13675*, 2022.

Laidlaw, C., Russell, S., and Dragan, A. Bridging rl theory and practice with the effective horizon. *arXiv preprint arXiv:2304.09853*, 2023.

Le, H., Wang, Y., Gotmare, A. D., Savarese, S., and Hoi, S. C. H. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 21314–21328, 2022.

Lee, K., Smith, L. M., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:235377145.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *ArXiv*, abs/2302.12192, 2023a. URL https://api.semanticscholar.org/CorpusID:257102772.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023b.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Lillicrap, T., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous Control with Deep Reinforcement Learning. *CoRR*, abs/1509.02971, 2016.

Lin, K., Li, D., He, X., Zhang, Z., and Sun, M.-T. Adversarial ranking for language generation. *Advances in neural information processing systems*, 30, 2017.

Liu, H., Trott, A., Socher, R., and Xiong, C. Competitive experience replay. In *International Conference on Learning Representations*, 2019.

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

Lu, X., Welleck, S., Jiang, L., Hessel, J., Qin, L., West, P., Ammanabrolu, P., and Choi, Y. Quark: Controllable text generation with reinforced unlearning. *arXiv preprint arXiv:2205.13636*, 2022.

Marbach, P. and Tsitsiklis, J. N. Approximate gradient methods in policy-space optimization of markov reward processes. *Discrete Event Dynamic Systems*, 13:111–148, 2003.

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Paulus, R., Xiong, C., and Socher, R. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Peters, J., Mulling, K., and Altun, Y. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010.

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. *arXiv preprint arXiv:2307.01952*, 2023.

Prabhudesai, M., Goyal, A., Pathak, D., and Fragkiadaki, K. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Ramachandran, G. S., Hashimoto, K., and Xiong, C. Causal-aware safe policy improvement for task-oriented dialogue. *arXiv preprint arXiv:2103.06370*, 2021.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Russell, S. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, 1998.

Ryang, S. and Abekawa, T. Framework of automatic text summarization using reinforcement learning. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 256–265, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1024.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. Trust Region Policy Optimization. *ArXiv*, abs/1502.05477, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Segalis, E., Valevski, D., Lumen, D., Matias, Y., and Leviathan, Y. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023.

Shi, Z., Chen, X., Qiu, X., and Huang, X. Toward diverse text generation with inverse reinforcement learning. *arXiv preprint arXiv:1804.11258*, 2018.

Shin, D., Brown, D. S., and Dragan, A. D. Offline preference-based apprenticeship learning. *arXiv preprint arXiv:2107.09251*, 2021.

Shu, R., Yoo, K. M., and Ha, J.-W. Reward optimization for neural machine translation with learned metrics. *arXiv preprint arXiv:2104.07541*, 2021.

Snell, C., Kostrikov, I., Su, Y., Yang, M., and Levine, S. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Takanobu, R., Zhu, H., and Huang, M. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. *arXiv preprint arXiv:1908.10719*, 2019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

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*, 2023a.

Wallace, B., Gokul, A., Ermon, S., and Naik, N. V. End-to-end diffusion latent optimization improves classifier guidance. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7246–7256, 2023b. URL https://api.semanticscholar.org/CorpusID:257757144.

Wang, B. and Vastola, J. J. Diffusion models generate images like painters: an analytical theory of outline first, details later, 2023.

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*, 2023a.

Wu, X., Sun, K., Zhu, F., Zhao, R., and Li, H. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023b.

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, S., Feng, Y., Zhang, S., and Zhou, M. Regularizing a model-based policy stationary distribution to stabilize offline reinforcement learning. In *International Conference on Machine Learning*, pp. 24980–25006. PMLR, 2022a.

Yang, S., Wang, Z., Zheng, H., Feng, Y., and Zhou, M. A regularized implicit policy for offline reinforcement learning. *arXiv preprint arXiv:2202.09673*, 2022b.

Yang, S., Zhang, S., Feng, Y., and Zhou, M. A unified framework for alternating offline model training and policy learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022c.

Yang, S., Zhang, S., Xia, C., Feng, Y., Xiong, C., and Zhou, M. Preference-grounded token-level guidance for language model fine-tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=6SRE9GZ9s6.

Yang, Z., Hu, Z., Dyer, C., Xing, E. P., and Berg-Kirkpatrick, T. Unsupervised text style transfer using language models as discriminators. *Advances in Neural Information Processing Systems*, 31, 2018.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Yuan, H., Yuan, Z., Tan, C., Wang, W., Huang, S., and Huang, F. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EdIGMCHk4l.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Ziebart, B. D. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pp. 1433–1438, 2008.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Appendix

Contents

1 Introduction **1**

2 Main Method **2**

3 Related Work **4**

4 Experiments **5**

5 Conclusion **9**

A Tabular Results **16**

B Detailed Method Derivation and Proofs **17**

C The Smaller γ, The Better? **24**

D Discussion on Our Method's Applicability to Real Human Preference **24**

E More Related Works **25**

F Experiment Details **27**

G More Generated Images **28**

A. Tabular Results

Table 4: HPSv2 and Aesthetic score for the multiple prompt experiment in Section 4.2. The first four result columns are the four styles in the HPSv2 test set and "Average" is the overall average. "Trajectory-level Reward" is the classical DPO-style objective discussed in Section 2.4, which assumes a latent trajectory-level reward function evaluating the entire T2I reverse chain as a whole. All baselines benchmarked in the HPSv2 paper are directly cited from the official Github Repository. Our produced results follow the testing principle in the HPSv2 paper and GitHub Repository. We bold the best result in each metric. Note that the HPSv2 paper and Github Repository do not report the Aesthetic score.

Model	Animation	Concept-art	Painting	Photo	Average	Aesthetic
GLIDE	23.34	23.08	23.27	24.50	23.55	-
LAFITE	24.63	24.38	24.43	25.81	24.81	-
VQ-Diffusion	24.97	24.70	25.01	25.71	25.10	-
FuseDream	25.26	25.15	25.13	25.57	25.28	-
Latent Diffusion	25.73	25.15	25.25	26.97	25.78	-
DALL·E mini	26.10	25.56	25.56	26.12	25.83	-
VQGAN + CLIP	26.44	26.53	26.47	26.12	26.39	-
CogView2	26.50	26.59	26.33	26.44	26.47	-
Versatile Diffusion	26.59	26.28	26.43	27.05	26.59	-
DALL·E 2	27.34	26.54	26.68	27.24	26.95	-
Stable Diffusion v1.4	27.26	26.61	26.66	27.27	26.95	-
Stable Diffusion v1.5	27.43	26.71	26.73	27.62	27.12	5.62
Stable Diffusion v2.0	27.48	26.89	26.86	27.46	27.17	-
Epic Diffusion	27.57	26.96	27.03	27.49	27.26	-
DeepFloyd-XL	27.64	26.83	26.86	27.75	27.27	-
Openjourney	27.85	27.18	27.25	27.53	27.45	-
MajicMix Realistic	27.88	27.19	27.22	27.64	27.48	-
ChilloutMix	27.92	27.29	27.32	27.61	27.54	-
Deliberate	28.13	27.46	27.45	27.62	27.67	-
SDXL Base 0.9	28.42	27.63	27.60	27.29	27.73	-
Realistic Vision	28.22	27.53	27.56	27.75	27.77	-
SDXL Refiner 0.9	28.45	27.66	27.67	27.46	27.80	-
Dreamlike Photoreal 2.0	28.24	27.60	27.59	27.99	27.86	-
Trajectory-level Reward	29.37	28.81	28.83	29.16	29.04	5.94
Ours	**30.46**	**29.95**	**30.01**	**29.93**	**30.09**	**6.31**

Table 5: HPSv2 and Aesthetic score for the ablation study on KL coefficient C in Section 4.3 **(c)**. Shown here are breakdown scores of our main method ($\gamma = 0.9$) in the *multiple* prompt experiment under various value of C, together with the best baseline in Table 4 of Appendix A. The first four result columns are the four styles in HPSv2 test set and "Average" is the overall average. Within subscript is one standard deviation, as plotted in Fig. 7b, calculated by the principle described in the HPSv2 paper and GitHub Repository.

Model	HPSv2					Aesthetic
	Animation	Concept-art	Painting	Photo	Averaged	
Baseline	29.37	28.81	28.83	29.16	29.04	5.94
$C = 7.5$	30.16	29.59	29.64	29.76	29.79 (0.26)	6.24
$C = 10.0$	30.36	29.87	29.91	29.80	29.99 (0.23)	6.29
$C = 12.5$	30.46	29.95	30.01	29.93	30.09 (0.22)	6.31
$C = 15.0$	30.30	29.72	29.74	29.77	29.88 (0.20)	6.23
$C = 20.0$	30.28	29.73	29.76	29.95	29.93 (0.20)	6.17

B. Detailed Method Derivation and Proofs

In this section, we provide a detailed step-by-step derivation of our method. For completeness and better readability, some materials in Section 2 will be restated.

B.1. Notation and Assumptions

This section restates the notations and assumptions in Section 2.1 for convenience.

Assumption 2.1. There is a latent reward function $r(s_t, a_t)$ that can score each step t of the T2I reverse chain.

We adopt the notations in prior works (*e.g.*, Fan et al., 2023; Black et al., 2023) to formulate the diffusion reverse process under the conditional generation setting as an Markov decision process (MDP), specified by $\mathcal{M} = (\mathbb{S}, \mathbb{A}, \mathcal{P}, r, \gamma, \rho)$. Specifically, let π_θ be the T2I with trainable parameters θ, *i.e.*, the policy network; $\{\boldsymbol{x}_t\}_{t=T}^{0}$ be the diffusion reverse chain of length T; and $\boldsymbol{c}$ be the conditioning variable, *i.e.*, the text conditional in our setting. We have, $\forall\, t$,

$$
\begin{gathered}
s_t \triangleq (\boldsymbol{x}_t, t, \boldsymbol{c})\,, \qquad a_t \triangleq \boldsymbol{x}_{t-1}\,, \qquad \pi_\theta(a_t \,|\, s_t) \triangleq p_\theta(\boldsymbol{x}_{t-1} \,|\, \boldsymbol{x}_t, t, \boldsymbol{c})\,, \\
\mathcal{P}(s_{t+1} \,|\, s_t, a_t) \triangleq \delta(\boldsymbol{x}_{t-1}, t-1, \boldsymbol{c})\,, \quad \rho(s_0) \triangleq (\mathcal{N}(\mathbf{0}, \boldsymbol{I}), \delta(T), \delta(\boldsymbol{c}))\,, \qquad r(s_t, a_t)\,,\ \gamma \in [0, 1]\,,
\end{gathered}
$$

where $\delta(\cdot)$ is the delta measure and $\mathcal{P}(s_{t+1} \,|\, s_t, a_t)$ is a deterministic transition. We denote the reverse chain generated by a (generic) T2I under the text conditional $\boldsymbol{c}$ as a trajectory τ, *i.e.*, $\tau \triangleq (s_0, a_0, s_1, a_1, \ldots, s_T) \iff (\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_0) \,|\, \boldsymbol{c}$. Note that for notation simplicity, $\boldsymbol{c}$ is absorbed into the state part of trajectory τ.

Similar to Wallace et al. (2023a), in the method derivation, we consider the setting where we are given two diffusion reverse chains (trajectories) with equal length T. For presentation simplicity, assume that τ^1 is the better trajectory, *i.e.*, $\tau^1 \succ \tau^2$. Let tuple $\mathrm{ord} \triangleq (1, 2)$ and $\sigma(\cdot)$ denotes the sigmoid function, *i.e.*, $\sigma(x) = \frac{1}{1+\exp(-x)}$.

Since in practice the state space of the T2I reverse chain is the continuous embedding space, it is self-evident to assume that any two trajectories do not cross with each other, as follows.

Assumption B.1 (No Crossing Trajectories)**.** $\forall\, \tau^i \neq \tau^j, s_t^i \neq s_t^j, \forall\, t \in \{0, \ldots, T\}$.

Furthermore, as in the standard RL setting (Sutton & Barto, 2018; Yang et al., 2022b), the reward function r in $\mathcal{M}$ needs to be bounded. Without loss of generality, we assume $r(s, a) \in [0, 1]$, and thus $r(s, a)$ may be interpreted as the probability of satisfying the preference when taking action a at state s.

Assumption 2.2. In $\mathcal{M}$, $\forall\, (s, a) \in \mathbb{S} \times \mathbb{A}$, $0 \leq r(s, a) \leq 1$.

In RL problems, the performance of a (generic) policy π is typically evaluated by the expected cumulative discounted rewards (Sutton & Barto, 2018), which is defined as,

$$
\eta(\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t\, r(s_t, a_t) \,|\, s_0 \sim \rho(\cdot), a_t \sim \pi(\cdot \,|\, s_t), s_{t+1} \sim \mathcal{P}(\cdot \,|\, s_t, a_t), \forall\, t \geq 0\right]. \tag{10}
$$

Note that Eq. (10) above is an extended version of Eq. (1) in Section 2.1.

Assumption 2.3. Based on Eq. (1), we assume that for a (generation) trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$, its quality is evaluated by $e(\tau) \triangleq \sum_{t=0}^{T} \gamma^t\, r(s_t, a_t)$.

Remark 2.4 in Section 2.1 provides a discussion on the practical rationality of $e(\tau)$ in T2I's preference alignment.

B.2. Step-by-step Derivation of Our Method

B.2.1. Expression of $e(\tau)$

We can express $\eta(\pi)$ in Eq. (10) by the (discounted) stationary distribution $d_\pi(s)$ of the policy π, defined as $d_\pi(s) \propto \sum_{t=0}^{T} \gamma^t \Pr(s_t = s \mid \pi, \mathcal{P})$, up to a (positive) normalizing constant (Yang et al., 2022a;c). We have

$$\begin{aligned}
\eta(\pi) &= \mathbb{E}_{a_t \sim \pi, s_{t+1} \sim \mathcal{P}} \left[\sum_{t=0}^{T} \gamma^t \, r(s_t, a_t) \right] = \sum_{t=0}^{T} \int_{\mathbb{S}} \Pr(s_t = s \mid \pi, \mathcal{P}) \int_{\mathbb{A}} \pi(a \mid s) \, \gamma^t \, r(s, a) \, \mathrm{d}a \, \mathrm{d}s \\
&= \int_{\mathbb{S}} \sum_{t=0}^{T} \gamma^t \Pr(s_t = s \mid \pi, \mathcal{P}) \int_{\mathbb{A}} \pi(a \mid s) \, r(s, a) \, \mathrm{d}a \, \mathrm{d}s \\
&\propto \int_{\mathbb{S}} d_\pi(s) \int_{\mathbb{A}} \pi(a \mid s) \, r(s, a) \, \mathrm{d}a \, \mathrm{d}s = \mathbb{E}_{s \sim d_\pi(s)} \mathbb{E}_{a \sim \pi(a \mid s)} \left[r(s, a) \right] .
\end{aligned}$$

The goal of RL is to maximize the expected cumulative discounted rewards $\eta(\pi)$, which is unfortunately difficult due to the complicate relationship between $d_\pi(s)$ and π. We therefore optimize an off-policy approximation of $\eta(\pi)$ by employing an approximation approach common in prior RL works (*e.g.*, Kakade & Langford, 2002; Peters et al., 2010; Schulman et al., 2015; Abdolmaleki et al., 2018; Peng et al., 2019). Specifically, we change $d_\pi(s)$ to $d_{\pi_O}(s)$ for some "old" policy π_O, from which we generate the off-policy trajectories/data. We further add a KL regularization on π towards the initial pre-trained model π_I to avoid generating unnatural images. In sum, we arrive at the following constrained policy search problem

$$\begin{aligned}
\arg\max_{\pi} \quad & \mathbb{E}_{s \sim d_{\pi_O}(s)} \mathbb{E}_{a \sim \pi(a \mid s)} \left[r(s, a) \right] \\
\text{s.t.} \quad & D_{\mathrm{KL}} \left(\pi(\cdot \mid s) \,\|\, \pi_I(\cdot \mid s) \right) \leq \epsilon , \quad \forall \, s \in \mathbb{S} \\
& \int_{\mathbb{A}} \pi(a \mid s) \, \mathrm{d}a = 1, \quad \forall \, s \in \mathbb{S},
\end{aligned} \tag{11}$$

where π_O may be chosen as π_I or some saved policy checkpoint not far away from π_I.

Enforcing the pointwise KL-regularization in Eq. (11) is difficult, as in AWR (Peng et al., 2019), we change the pointwise KL-regularization into enforcing the regularization only in expectation $\mathbb{E}_{s \sim d_{\pi_O}} \left[\cdots \right]$ and change Eq. (11) into a regularized maximization problem

$$\begin{aligned}
\arg\max_{\pi} \quad & \mathbb{E}_{s \sim d_{\pi_O}(s)} \mathbb{E}_{a \sim \pi(a \mid s)} \left[r(s, a) \right] - C \cdot \mathbb{E}_{s \sim d_{\pi_O}(s)} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid s) \,\|\, \pi_I(\cdot \mid s) \right) \right] \\
\text{s.t.} \quad & \int_{\mathbb{A}} \pi(a \mid s) \, \mathrm{d}a = 1, \quad \forall \, s \in \mathbb{S}.
\end{aligned} \tag{12}$$

The Lagrange form of the maximization problem Eq. (12) is

$$\mathcal{L}(\pi) \triangleq \mathbb{E}_{s \sim d_{\pi_O}(s)} \mathbb{E}_{a \sim \pi(a \mid s)} \left[r(s, a) \right] - C \cdot \mathbb{E}_{s \sim d_{\pi_O}(s)} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid s) \,\|\, \pi_I(\cdot \mid s) \right) \right] + \int_{\mathbb{S}} \alpha_s \left(1 - \int_{\mathbb{A}} \pi(a \mid s) \, \mathrm{d}a \right) \mathrm{d}s \, . \tag{13}$$

$\forall \, s \in \mathbb{S}, a \in \mathbb{A}$, the optimal policy under $\mathcal{L}(\pi)$ can be obtained by setting the derivatives *w.r.t.* $\pi(a \mid s)$ equal to 0. We have

$$\begin{aligned}
& \frac{\partial \, \mathcal{L}(\pi)}{\partial \, \pi(a \mid s)} = d_{\pi_O}(s) r(s, a) - C d_{\pi_O}(s) \log \pi(a \mid s) - C d_{\pi_O}(s) + C d_{\pi_O}(s) \log \pi_I(a \mid s) - \alpha_s = 0 \\
\implies \quad & r(s, a) = C \log \frac{\pi^*(a \mid s)}{\pi_I(a \mid s)} + C + \frac{\alpha_s}{d_{\pi_O}(s)}
\end{aligned} \tag{14}$$

where π^* is the optimal policy under r.

From Eq. (14), we can also get the formula for the optimal policy π^* as

$$\pi^*(a \mid s) = \exp\left(\frac{1}{C} r(s, a) \right) \pi_I(a \mid s) \exp\left(-1 - \frac{\alpha_s}{C d_{\pi_O}(s)} \right) \triangleq \exp\left(\frac{1}{C} r(s, a) \right) \pi_I(a \mid s) \frac{1}{Z(s)}, \tag{15}$$

where $Z(s)$ denotes the partition function, taking the form

$$Z(s) = \int_{\mathbb{A}} \exp\left(\frac{1}{C} r(s,a)\right) \pi_I(a \,|\, s)\, \mathrm{d}a.$$

For a *given* trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$, the quality evaluation $e(\tau)$ can be expressed by π^* as

$$\begin{aligned} e(\tau) &\triangleq \sum_{t=0}^{T} \gamma^t\, r(s_t, a_t) = C \sum_{t=0}^{T} \gamma^t \log \frac{\pi^*(a_t \,|\, s_t)}{\pi_I(a_t \,|\, s_t)} + C \sum_{t=0}^{T} \gamma^t \left(1 + \frac{\alpha_{s_t}}{C d_{\pi_O}(s_t)}\right) \\ &= C \sum_{t=0}^{T} \gamma^t \log \frac{\pi^*(a_t \,|\, s_t)}{\pi_I(a_t \,|\, s_t)} + C \sum_{t=0}^{T} \gamma^t \log Z(s_t). \end{aligned} \tag{16}$$

Since the trajectory τ and hence all s_t's are given, $Z(s_t)$'s are constant and the summation over $\log Z(s_t)$ is a "property" of the trajectory τ, we thus denote $\log Z(\tau) \triangleq \sum_{t=0}^{T} \gamma^t \log Z(s_t)$ for notation simplicity. Then the formula for $e(\tau)$ becomes

$$e(\tau) = C \sum_{t=0}^{T} \left[\gamma^t \log \frac{\pi^*(a_t \,|\, s_t)}{\pi_I(a_t \,|\, s_t)}\right] + C \log Z(\tau)\,. \tag{17}$$

B.2.2. Loss Function for T2I/Policy Training

Recall that we are given two diffusion reverse chains (trajectories) $\left\{\tau^1, \tau^2\right\}$ with equal length T. Also recall the notation that τ^1 is the better trajectory, *i.e.*, $\tau^1 \succ \tau^2$, the tuple $\mathrm{ord} \triangleq (1, 2)$ and $\sigma(\cdot)$ denotes the sigmoid function. Under the Bradley-Terry model of pairwise preference, the probability of ord under $\left\{e\left(\tau^k\right)\right\}_{k=1}^2$ and hence π^* is

$$\Pr\left(\mathrm{ord} \,|\, \pi^*, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right) = \frac{\exp\left(e\left(\tau^1\right)\right)}{\exp\left(e\left(\tau^1\right)\right) + \exp\left(e\left(\tau^2\right)\right)} = \sigma\left(e\left(\tau^1\right) - e\left(\tau^2\right)\right)\,, \tag{18}$$

where we explicitly put π^* into the conditioning variables for better readability.

From Eq. (17), $e\left(\tau^1\right)$ and $e\left(\tau^2\right)$ respectively contains the "partition functions" $Z(\tau^1)$ and $Z(\tau^2)$, both of which are intractable. We argue that $Z(\tau^1) \geq Z(\tau^2)$, which will be critical for providing a tractable lower bound of Eq. (18) that cancels out these partition functions. Our argument is based on the reward-shaping technique (Ng et al., 1999), as follows.

Definition 2.5 (Reward Shaping)**.** A shaping-reward function Φ is a real-valued function on the state space, $\Phi : \mathbb{S} \to \mathbb{R}$. It induces a new MDP $\mathcal{M}' = (\mathbb{S}, \mathbb{A}, \mathcal{P}, r', \gamma, \rho)$ where $r'(s,a) \triangleq r(s,a) + \Phi(s)$.

Lemma 2.6 (Invariance of Optimal Policy under Reward Shaping)**.** *The optimal (regularized) policy Eq. (3) under the reward-shaped MDP $\mathcal{M}'$ is the same as that in the original MDP $\mathcal{M}$.*

Remark B.2. The only difference between the MDPs $\mathcal{M}$ and $\mathcal{M}'$ is the reward function (r *v.s.* r'). In particular, they share the same state and action space. Therefore, it make sense to consider the invariance of the optimal policy in these two MDPs. Invariance means that, in these two MDPs, at each state, the optimal policies take the same action with the same probability.

Proof of Lemma 2.6. Denote the optimal policy under the MDP $\mathcal{M}'$ as $\pi^{*'}$, we have

$$\begin{aligned} \pi^{*'}(a \,|\, s) &= \frac{\exp\left(\frac{1}{C}\left(r(s,a) + \Phi(s)\right)\right) \pi_I(a \,|\, s)}{\int_{\mathbb{A}} \exp\left(\frac{1}{C}\left(r(s,a) + \Phi(s)\right)\right) \pi_I(a \,|\, s)\, \mathrm{d}a} = \frac{\exp\left(\frac{1}{C}\Phi(s)\right) \exp\left(\frac{1}{C} r(s,a)\right) \pi_I(a \,|\, s)}{\exp\left(\frac{1}{C}\Phi(s)\right) \int_{\mathbb{A}} \exp\left(\frac{1}{C} r(s,a)\right) \pi_I(a \,|\, s)\, \mathrm{d}a} \\ &= \pi^*(a \,|\, s), \end{aligned} \tag{19}$$

since $\exp\left(\frac{1}{C}\Phi(s)\right)$ is independent of the integration dummy-variable a in the denominator. □

Definition 2.7. The equivalence class $[r]$ of the reward function r is the set of all reward functions that can be obtained from r by reward shaping, *i.e.*, $\forall\, r' \in [r], \exists\, \Phi : \mathbb{S} \to \mathbb{R},\ s.t.\ r'(s,a) - r(s,a) = \Phi(s), \forall\, s \in \mathbb{S}, a \in \mathbb{A}$.

Remark 2.8. By Lemma 2.6, all reward functions in $[r]$ share the same optimal (regularized) policy as r, *i.e.*, Eq. (3).

With the reshaping technique, we can justify our previous argument that $Z(\tau^1) \geq Z(\tau^2)$ as follows.

Theorem 2.9. *Under Assumption 2.2, and a sufficiently large regularization coefficient C, for any finite number $K \geq 2$ of trajectories $\{\tau^k\}_{k=1}^K$ where $\tau^1 \succ \tau^2 \succ \cdots \succ \tau^K$, $\exists\, r' \in [r]$, s.t., $Z(\tau^1) \geq Z(\tau^2) \geq \cdots \geq Z(\tau^K)$ under r'.*

We defer the proof of Theorem 2.9 to Section B.3.

Remark 2.10. For the value of C, as we will see in the proof, we technically require that $\forall (s,a) \in \mathbb{S} \times \mathbb{A}, r(s,a)/C \leq \text{const} \approx 1.79$. Under Assumption 2.2, $C \geq 0.56$ will suffice. We note that this technical requirement helps reducing the search space of the hyperparameter C in practice.

Remark 2.11. By Lemma 2.6, r' in Theorem 2.9 and the original r lead to the same optimal policy π^*, which is our ultimate target. Due to this invariance, for notation simplicity, we hereafter refer to r' as r, though we *may* actually work in the "equivalent" MDP $\mathcal{M}' = (\mathbb{S}, \mathbb{A}, \mathcal{P}, r', \gamma, \rho)$.

With Theorem 2.9, we can lower bound $\Pr\left(\text{ord} \,|\, \pi^*, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)$ in Eq. (18) by a simpler formula. After plugging the expression of $e(\tau)$ *w.r.t.* the optimal policy π^* in Eq. (17), we have,

$$
\begin{aligned}
\Pr\left(\text{ord} \,|\, \pi^*, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right) &= \frac{\exp\left(C \sum_{t=0}^T \gamma^t \log \frac{\pi^*\left(a_t^1 \,|\, s_t^1\right)}{\pi_I\left(a_t^1 \,|\, s_t^1\right)}\right) Z\left(\tau^1\right)^C}{\sum_{i=1}^2 \exp\left(C \sum_{t=0}^T \gamma^t \log \frac{\pi^*\left(a_t^i \,|\, s_t^i\right)}{\pi_I\left(a_t^i \,|\, s_t^i\right)}\right) Z\left(\tau^i\right)^C} \\
&\geq \frac{\exp\left(C \sum_{t=0}^T \gamma^t \log \frac{\pi^*\left(a_t^1 \,|\, s_t^1\right)}{\pi_I\left(a_t^1 \,|\, s_t^1\right)}\right) Z\left(\tau^1\right)^C}{\sum_{i=1}^2 \exp\left(C \sum_{t=0}^T \gamma^t \log \frac{\pi^*\left(a_t^i \,|\, s_t^i\right)}{\pi_I\left(a_t^i \,|\, s_t^i\right)}\right) Z\left(\tau^1\right)^C} \\
&= \frac{\exp\left(C \sum_{t=0}^T \gamma^t \log \frac{\pi^*\left(a_t^1 \,|\, s_t^1\right)}{\pi_I\left(a_t^1 \,|\, s_t^1\right)}\right)}{\sum_{i=1}^2 \exp\left(C \sum_{t=0}^T \gamma^t \log \frac{\pi^*\left(a_t^i \,|\, s_t^i\right)}{\pi_I\left(a_t^i \,|\, s_t^i\right)}\right)} .
\end{aligned}
\tag{20}
$$

By our definition on the quality evaluation $e(\tau)$, a better trajectory τ comes with a higher $e(\tau)$. Hence $\exp\left(e\left(\tau^1\right)\right) / \left(\sum_{i=1}^2 \exp\left(e\left(\tau^i\right)\right)\right) \geq \exp\left(e\left(\tau^2\right)\right) / \left(\sum_{i=1}^2 \exp\left(e\left(\tau^i\right)\right)\right)$. In other words, among $\{\tau^1, \tau^2\}$, τ^1 should have the highest chance of being ranked top under the Bradley-Terry preference model Eq. (18) induced by the true reward $r(s,a)$. Thus we conclude that $\Pr\left(\text{ord} \,|\, \pi^*, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right) = \max \Pr\left(\cdot \,|\, \pi^*, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)$, *i.e.*, in the MDP $\mathcal{M}$ (or $\mathcal{M}'$) with the addition of (π_I, C) and conditioning on π^*, ord should be the most probable ordering under the Bradley-Terry model Eq. (18). Thus, in order to approximate π^*, our parametrized policy π_θ ought to maximize the likelihood of ord under the *corresponding Bradley-Terry model* constructed by substituting π^* with π_θ. Based on this intuition, we train π_θ by maximizing the lower bound of the Bradley-Terry likelihood of ord in Eq. (20), which leads to the negative-log-likelihood objective for an minimization problem for training π_θ as

$$
\begin{aligned}
\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right) &= -\log \sigma\left(C \sum_{t=0}^T \gamma^t \left[\log \frac{\pi_\theta\left(a_t^1 \,|\, s_t^1\right)}{\pi_I\left(a_t^1 \,|\, s_t^1\right)} - \log \frac{\pi_\theta\left(a_t^2 \,|\, s_t^2\right)}{\pi_I\left(a_t^2 \,|\, s_t^2\right)}\right]\right) \\
&= -\log \sigma\left(C \times \frac{1-\gamma^{T+1}}{1-\gamma} \mathbb{E}_{t \sim \text{Cat}(\{\gamma^t\})}\left[\log \frac{\pi_\theta(a_t^1 \mid s_t^1)}{\pi_I(a_t^1 \mid s_t^1)} - \log \frac{\pi_\theta(a_t^2 \mid s_t^2)}{\pi_I(a_t^2 \mid s_t^2)}\right]\right) \\
&= -\log \sigma\left(\left(C \times \frac{1-\gamma^{T+1}}{1-\gamma}\right) \mathbb{E}_{t \sim \text{Cat}(\{\gamma^t\})}\left[\log \frac{\pi_\theta(a_t^1 \mid s_t^1)}{\pi_I(a_t^1 \mid s_t^1)} - \log \frac{\pi_\theta(a_t^2 \mid s_t^2)}{\pi_I(a_t^2 \mid s_t^2)}\right]\right) \\
&= -\log \sigma\left(C\, \mathbb{E}_{t \sim \text{Cat}(\{\gamma^t\})}\left[\log \frac{\pi_\theta(a_t^1 \mid s_t^1)}{\pi_I(a_t^1 \mid s_t^1)} - \log \frac{\pi_\theta(a_t^2 \mid s_t^2)}{\pi_I(a_t^2 \mid s_t^2)}\right]\right) \\
&\text{with} \quad C \leftarrow C \times \frac{1-\gamma^{T+1}}{1-\gamma} ,
\end{aligned}
\tag{21}
$$

where $\text{Cat}\left(\{\gamma^t\}\right)$ denotes the categorical distribution on $\{0, \ldots, T\}$ with the probability vector $\{\gamma^t / \sum_{t'} \gamma^{t'}\}_{t=0}^T$; and C is overloaded to absorb the normalization constant, which is legitimated given that C itself is a hyperparameter and so does C times the normalization constant.

B.3. Proofs

B.3.1. Derivation of the Gradient in Eq. (9)

Here we derive the gradient of $\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)$ in Eq. (21) with respect to θ, which is presented in Section 2.2.

Since Eq. (21) is an objective for a minimization problem, the gradient update direction is $-\nabla_\theta \mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right) = \nabla_\theta\left(-\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)\right)$. The gradient can be derived by chain rule as follows. For notation simplicity, we denote $\widetilde{e}(\tau^k) \triangleq C\sum_{t=0}^T \gamma^t \log \frac{\pi_\theta\left(a_t^k \,|\, s_t^k\right)}{\pi_I\left(a_t^k \,|\, s_t^k\right)}$. We have

$$-\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right) = -\log\left(1+\exp\left(\widetilde{e}\left(\tau^2\right)-\widetilde{e}\left(\tau^1\right)\right)\right)$$

$$\frac{\partial\left(-\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)\right)}{\partial\left(\widetilde{e}\left(\tau^2\right)-\widetilde{e}\left(\tau^1\right)\right)} = -\frac{\exp\left(\widetilde{e}\left(\tau^2\right)\right)}{\exp\left(\widetilde{e}\left(\tau^1\right)\right)+\exp\left(\widetilde{e}\left(\tau^2\right)\right)}$$

$$\begin{aligned}\forall\, k=1,2, \quad \frac{\partial\,\widetilde{e}\left(\tau^k\right)}{\partial\,\theta} &= C\sum_{t=0}^T \gamma^t \nabla_\theta \log\frac{\pi_\theta\left(a_t^k \,|\, s_t^k\right)}{\pi_I\left(a_t^k \,|\, s_t^k\right)} = C\sum_{t=0}^T \gamma^t \frac{\pi_I\left(a_t^k \,|\, s_t^k\right)}{\pi_\theta\left(a_t^k \,|\, s_t^k\right)} \nabla_\theta \pi_\theta\left(a_t^k \,|\, s_t^k\right) \\ &= C\sum_{t=0}^T \gamma^t \pi_I\left(a_t^k \,|\, s_t^k\right) \nabla_\theta \log \pi_\theta\left(a_t^k \,|\, s_t^k\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial\left(-\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)\right)}{\partial\,\theta} &= \frac{\partial\left(-\mathcal{L}_\gamma\left(\theta \,|\, \text{ord}, \left\{e\left(\tau^k\right)\right\}_{k=1}^2\right)\right)}{\partial\left(\widetilde{e}\left(\tau^2\right)-\widetilde{e}\left(\tau^1\right)\right)}\left(\frac{\partial\,\widetilde{e}\left(\tau^2\right)}{\partial\,\theta}-\frac{\partial\,\widetilde{e}\left(\tau^1\right)}{\partial\,\theta}\right) \\ &= -\frac{\exp\left(\widetilde{e}\left(\tau^2\right)\right)}{\exp\left(\widetilde{e}\left(\tau^1\right)\right)+\exp\left(\widetilde{e}\left(\tau^2\right)\right)}\left(\frac{\partial\,\widetilde{e}\left(\tau^2\right)}{\partial\,\theta}-\frac{\partial\,\widetilde{e}\left(\tau^1\right)}{\partial\,\theta}\right) \\ &= \frac{\exp\left(\widetilde{e}\left(\tau^2\right)\right)}{\exp\left(\widetilde{e}\left(\tau^1\right)\right)+\exp\left(\widetilde{e}\left(\tau^2\right)\right)}\left(\frac{\partial\,\widetilde{e}\left(\tau^1\right)}{\partial\,\theta}-\frac{\partial\,\widetilde{e}\left(\tau^2\right)}{\partial\,\theta}\right) \\ &= \frac{\exp\left(\widetilde{e}\left(\tau^2\right)-\widetilde{e}\left(\tau^1\right)\right)}{1+\exp\left(\widetilde{e}\left(\tau^2\right)-\widetilde{e}\left(\tau^1\right)\right)} \times C \times \sum_{t=0}^T \gamma^t \Bigg(\pi_I\left(a_t^1 \,|\, s_t^1\right)\nabla_\theta \log\pi_\theta\left(a_t^1 \,|\, s_t^1\right) \\ &\qquad - \pi_I\left(a_t^2 \,|\, s_t^2\right)\nabla_\theta \log\pi_\theta\left(a_t^2 \,|\, s_t^2\right)\Bigg).\end{aligned}$$

B.3.2. Proof of Theorem 2.9

As a reminder, in Theorem 2.9 we consider a more general case where we are given a finite number K of trajectories whose preference ordering is assume to be $\tau^1 \succ \tau^2 \succ \cdots \succ \tau^K$. Each trajectory τ^k takes the form $\tau^k = (s_0^k, a_0^k, s_1^k, a_1^k, \ldots, s_T^k)$.

A Simplified Case without Reward Shaping.

To gain some intuitions, we first present a simplified setting where the distribution π_I is deterministic on the given samples, *i.e.*, $\pi_I\left(a_t^i \,|\, s_t^i\right) = \delta\left(a_t^i \,|\, s_t^i\right)$. In this scenario, Theorem 2.9 can be proved without using the reward-shaping argument.

We now state and proof this special case of Theorem 2.9.

Theorem B.3 (A special case of Theorem 2.9)**.** *If the sampling distribution $\pi_I\left(a_t^i \,|\, s_t^i\right) = \delta\left(a_t^i \,|\, s_t^i\right)$, then the original reward function $r(s,a)$ satisfies $Z(\tau^1) \geq Z(\tau^2) \geq \cdots \geq Z(\tau^K)$.*

Proof. Our target is $\forall\, k \in \{1, \ldots, K\}, i \in \{k, \ldots, K\}$,

$$Z(\tau^k) \geq Z(\tau^i) \iff \log Z(\tau^k) \geq \log Z(\tau^i) \iff \sum_{t=0}^{T} \gamma^t \log Z\left(s_t^k\right) \geq \sum_{t=0}^{T} \gamma^t \log Z\left(s_t^i\right)$$
$$\iff \sum_{t=0}^{T} \gamma^t \left(\log Z\left(s_t^k\right) - \log Z\left(s_t^i\right)\right) \geq 0.$$

In the special case of $\pi_I\left(a_t^i \,|\, s_t^i\right) = \delta\left(a_t^i \,|\, s_t^i\right)$, with the *original* reward function $r(s, a)$, we have

$$Z\left(s_t^i\right) = \int_{\mathbb{A}} \exp\left(\frac{1}{C} r\left(s_t^i, a\right)\right) \pi_I\left(a \,|\, s_t^i\right) \mathrm{d}a = \exp\left(\frac{1}{C} r\left(s_t^i, a_t^i\right)\right) \implies \log Z\left(s_t^i\right) = \frac{1}{C} r\left(s_t^i, a_t^i\right)$$
$$\implies \log Z\left(s_t^k\right) - \log Z\left(s_t^i\right) = \frac{1}{C}\left(r\left(s_t^k, a_t^k\right) - r\left(s_t^i, a_t^i\right)\right)$$
$$\implies \sum_{t=0}^{T} \gamma^t \left(\log Z\left(s_t^k\right) - \log Z\left(s_t^i\right)\right) = \frac{1}{C} \sum_{t=0}^{T} \gamma^t \left(r\left(s_t^k, a_t^k\right) - r\left(s_t^i, a_t^i\right)\right) .$$

Since $\tau^k \succ \tau^i \iff e\left(\tau^k\right) > e\left(\tau^i\right)$, plugging in the definition of $e(\tau)$, we get,

$$e\left(\tau^k\right) > e\left(\tau^i\right) \iff \sum_{t=0}^{T} \gamma^t r(s_t^k, a_t^k) \geq \sum_{t=0}^{T} \gamma^t r(s_t^i, a_t^i) \iff \sum_{t=0}^{T} \gamma^t \left(r(s_t^k, a_t^k) - r(s_t^i, a_t^i)\right) \geq 0 \iff$$
$$\sum_{t=0}^{T} \gamma^t \log Z\left(s_t^k\right) \geq \sum_{t=0}^{T} \gamma^t \log Z\left(s_t^i\right) \iff \log Z\left(\tau^k\right) \geq \log Z\left(\tau^i\right) \iff Z\left(\tau^k\right) \geq Z\left(\tau^i\right) .$$

Hence the original reward function $r(s, a)$, *without shaping*, satisfies the ordering $Z\left(\tau^k\right) \geq Z\left(\tau^i\right)$. Notice that all the above steps are " $\iff$ " and recall our assumption that $\tau^1 \succ \tau^2 \succ \cdots \succ \tau^K \implies \tau^k \succ \tau^i \iff k \leq i$. It is clear that such an ordering is transitive, in a sense that, if $\tau^k \succ \tau^i \succ \tau^j$, then

$$\left.\begin{array}{l} \tau^k \succ \tau^i \implies k \leq i \\ \tau^i \succ \tau^j \implies i \leq j \end{array}\right\} \implies k \leq j \implies Z\left(\tau^k\right) \geq Z\left(\tau^j\right) .$$

Since k is arbitrary, we conclude that $Z(\tau^1) \geq Z(\tau^2) \geq \cdots \geq Z(\tau^K)$, as desired. □

The General Case

We repead Theorem 2.9 here for better readability.

Theorem 2.9. *Under Assumption 2.2, and a sufficiently large regularization coefficient C, for any finite number $K \geq 2$ of trajectories $\{\tau^k\}_{k=1}^K$ where $\tau^1 \succ \tau^2 \succ \cdots \succ \tau^K$, $\exists\, r' \in [r]$, s.t., $Z(\tau^1) \geq Z(\tau^2) \geq \cdots \geq Z(\tau^K)$ under r'.*

As discussed in Remark 2.10, for the value of C, we technically requires that $\forall\,(s, a) \in \mathbb{S} \times \mathbb{A}, r(s, a)/C \leq \text{const} \approx 1.79$. Hence, under Assumption 2.2, $C \geq 0.56$ will suffice. This provides some information on the setting of hyperparameter C.

Proof. Under the shaped reward $r'(s, a) = r(s, a) + \Phi(s)$, $Z\left(s_t^k\right)$ takes the form

$$\begin{aligned} Z\left(s_t^k\right) &= \int_{\mathbb{A}} \exp\left(\frac{1}{C} r\left(s_t^k, a\right) + \frac{1}{C} \Phi\left(s_t^k\right)\right) \pi_I\left(a \,|\, s_t^k\right) \mathrm{d}a \\ &= \exp\left(\frac{1}{C} \Phi\left(s_t^k\right)\right) \int_{\mathbb{A}} \exp\left(\frac{1}{C} r\left(s_t^k, a\right)\right) \pi_I\left(a \,|\, s_t^k\right) \mathrm{d}a \\ &= \exp\left(\frac{1}{C} \Phi\left(s_t^k\right)\right) \mathbb{E}_{a \sim \pi_I\left(\cdot \,|\, s_t^k\right)} \left[\exp\left(\frac{1}{C} r\left(s_t^k, a\right)\right)\right] . \end{aligned}$$

Taking $\log$ on both sides of the equation, by Jensen's inequality, we have

$$\log Z\left(s_t^k\right)=\frac{1}{C}\Phi\left(s_t^k\right)+\log\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^k\right)}\left[\exp\left(\frac{1}{C}r\left(s_t^k,a\right)\right)\right]\geq\frac{1}{C}\Phi\left(s_t^k\right)+\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^k\right)}\left[\frac{1}{C}r\left(s_t^k,a\right)\right].$$

On the other hand, we also have

$$Z\left(s_t^i\right)=\exp\left(\frac{1}{C}\Phi\left(s_t^i\right)\right)\int_{\mathbb{A}}\exp\left(\frac{1}{C}r\left(s_t^i,a\right)\right)\pi_I\left(a\,|\,s_t^i\right)\mathrm{d}a=\exp\left(\frac{1}{C}\Phi\left(s_t^i\right)\right)\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[\exp\left(\frac{1}{C}r\left(s_t^i,a\right)\right)\right],$$

Taking $\log$ again on both sides of the equations, we have

$$\begin{aligned}\log Z\left(s_t^i\right)&=\frac{1}{C}\Phi\left(s_t^i\right)+\log\left(\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[\exp\left(\frac{1}{C}r\left(s_t^i,a\right)\right)\right]\right)\\&\leq\frac{1}{C}\Phi\left(s_t^i\right)+\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[\exp\left(\frac{1}{C}r\left(s_t^i,a\right)\right)\right]-1\\&\leq\frac{1}{C}\Phi\left(s_t^i\right)+\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[1+\frac{1}{C}r\left(s_t^i,a\right)+\frac{1}{C^2}r^2\left(s_t^i,a\right)\right]-1\\&\leq\frac{1}{C}\Phi\left(s_t^i\right)+\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[\frac{1}{C}r\left(s_t^i,a\right)+\frac{1}{C^2}r\left(s_t^i,a\right)\right]\\&=\frac{1}{C}\Phi\left(s_t^i\right)+\frac{C+1}{C^2}\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[r\left(s_t^i,a\right)\right]\end{aligned}$$

where the first inequality is because $\forall\, x>0,\log x\leq x-1$; the second inequality is because $e^x\leq 1+x+x^2,\forall\, x<\mathrm{const}\approx 1.79$; the third inequality is because Assumption 2.2, *i.e.*, $\forall\,(s,a)\in\mathbb{S}\times\mathbb{A},\,0\leq r(s,a)\leq 1$.

Combining the above analysis, we have

$$\begin{aligned}\log Z\left(s_t^k\right)-\log Z\left(s_t^i\right)&\geq\frac{1}{C}\Phi\left(s_t^k\right)+\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^k\right)}\left[\frac{1}{C}r\left(s_t^k,a\right)\right]-\frac{1}{C}\Phi\left(s_t^i\right)-\frac{C+1}{C^2}\mathbb{E}_{a\sim\pi_I\left(\cdot\,|\,s_t^i\right)}\left[r\left(s_t^i,a\right)\right]\\&\geq\frac{1}{C}\left(\Phi\left(s_t^k\right)-\Phi\left(s_t^i\right)\right)-\frac{C+1}{C^2}\end{aligned}$$

where the second inequality is again due to Assumption 2.2, *i.e.*, $0\leq r(s,a)\leq 1$.

Summing over t, we have

$$\sum_{t=0}^{T}\gamma^t\left(\log Z\left(s_t^k\right)-\log Z\left(s_t^i\right)\right)\geq\frac{1}{C}\sum_{t=0}^{T}\gamma^t\left(\Phi\left(s_t^k\right)-\Phi\left(s_t^i\right)\right)-\frac{C+1}{C^2}\frac{1-\gamma^{T+1}}{1-\gamma}\geq_?0\,,$$

where the desired final inequality of ≥ 0 holds if

$$\sum_{t=0}^{T}\gamma^t\left(\Phi\left(s_t^k\right)-\Phi\left(s_t^i\right)\right)\geq\frac{C+1}{C}\frac{1-\gamma^{T+1}}{1-\gamma}\,,$$

where $\frac{C+1}{C}\frac{1-\gamma^{T+1}}{1-\gamma}<\infty$ is finite. Therefore, there exists a finite shaping function $\Phi(s)$ satisfying the above constraint, which can restore the order of $Z(\tau^k)$ and $Z(\tau^i)$ to be $Z(\tau^k)\geq Z(\tau^i)$.

Furthermore, this restoration is transitive in the sense that, for $\tau^k\succ\tau^i\succ\tau^j$ and the corresponding $Z(\tau^k)$, $Z(\tau^i)$, and $Z(\tau^j)$, if

$$\begin{aligned}\textstyle\sum_{t=0}^{T}\gamma^t\left(\Phi\left(s_t^k\right)-\Phi\left(s_t^i\right)\right)\geq\frac{C+1}{C}\frac{1-\gamma^{T+1}}{1-\gamma}&\implies Z(\tau^k)\geq Z(\tau^i)\\\textstyle\sum_{t=0}^{T}\gamma^t\left(\Phi\left(s_t^i\right)-\Phi\left(s_t^j\right)\right)\geq\frac{C+1}{C}\frac{1-\gamma^{T+1}}{1-\gamma}&\implies Z(\tau^i)\geq Z(\tau^j),\end{aligned}$$

then $\log Z(\tau^k) - \log Z(\tau^j) \geq 0 \iff Z(\tau^k) \geq Z(\tau^j)$, because,

$$\begin{aligned}
\log Z(\tau^k) - \log Z(\tau^j) &= \sum_{t=0}^{T} \gamma^t \left(\log Z\left(s_t^k\right) - \log Z\left(s_t^j\right) \right) \\
&= \sum_{t=0}^{T} \gamma^t \left(\log Z\left(s_t^k\right) - \log Z\left(s_t^i\right) + \log Z\left(s_t^i\right) - \log Z\left(s_t^j\right) \right) \\
&= \sum_{t=0}^{T} \gamma^t \left(\log Z\left(s_t^k\right) - \log Z\left(s_t^i\right) \right) + \sum_{t=0}^{T} \gamma^t \left(\log Z\left(s_t^i\right) - \log Z\left(s_t^j\right) \right) \\
&\geq \frac{1}{C} \left(\sum_{t=0}^{T} \gamma^t \left(\Phi\left(s_t^k\right) - \Phi\left(s_t^i\right) \right) - \frac{C+1}{C} \frac{1-\gamma^{T+1}}{1-\gamma} \right) \\
&\quad + \frac{1}{C} \left(\sum_{t=0}^{T} \gamma^t \left(\Phi\left(s_t^i\right) - \Phi\left(s_t^j\right) \right) - \frac{C+1}{C} \frac{1-\gamma^{T+1}}{1-\gamma} \right) \geq 0 \,,
\end{aligned}$$

since by Assumption 2.2, $C \geq 0.56$ is positive.

It follows that for K trajectories $\tau^1 \succ \tau^2 \succ \cdots \succ \tau^K$, we can restore the order of $Z\left(\tau^k\right)$'s by **at most** $(K-1)$ requirements on the reward-shaping function $\Phi(s)$, taking the form,

$$\begin{aligned}
\sum_{t=0}^{T} \gamma^t \left(\Phi\left(s_t^1\right) - \Phi\left(s_t^2\right) \right) &\geq \frac{C+1}{C} \frac{1-\gamma^{T+1}}{1-\gamma}, \\
&\vdots \\
\sum_{t=0}^{T} \gamma^t \left(\Phi\left(s_t^{K-1}\right) - \Phi\left(s_t^K\right) \right) &\geq \frac{C+1}{C} \frac{1-\gamma^{T+1}}{1-\gamma} \,.
\end{aligned}$$

Since each of these $(K-1)$ requirements only specify a finite lower bound on the discounted sum of the difference of the reward-shaping function $\Phi(\cdot)$ on two trajectories, it is clear that there exists a finite reward-shaping function $\Phi(s)$ satisfying these requirements. Hence, there exists a shaped reward function $r' \in [r], r'(s,a) = r(s,a) + \Phi(s)$, such that $Z(\tau^1) \geq \cdots \geq Z(\tau^K)$ under r'. □

C. The Smaller γ, The Better?

Though it would be great if "the smaller γ, the better result", this is unfortunately not true. In the multiple prompt experiment, as shown in Fig. 6b, $\gamma = 0.95$ is slightly better than $\gamma = 0.9$ towards the end of training.

As another verification, we re-run our single prompt experiment ("A green colored rabbit.") under $\gamma = 0.8$. Fig. 8 compares its performance with $\gamma \in \{0.9, 1.0\}$ at each decile of the training process. From Fig. 8, we see that $\gamma = 0.8$ is again better than the classical setting of $\gamma = 1.0$ and indeed trains faster than $\gamma = 0.9$ in the first 20% of the training process. However, in the second half of training, $\gamma = 0.8$ is less stable and its performance is inferior to $\gamma = 0.9$.

Recall that during training, the smaller γ, the more emphasis is on the initial steps of the reverse chain. As shown in Fig. 8, a too-small γ may thus have a stronger tendency of overfitting, leading to a more varying training process and inferior final result. Further, during training, a too-small γ may pay too-few attention to the later steps of the reverse chain that generate image details, resulting in less preferable image generations.

From Figs. 6b and 8, we conclude that while a *sensible* incorporation of $\gamma < 1$ can outperform the classical setting of $\gamma = 1$, final performance is not monotone with γ. The optimal γ value can be task specific. In our experiments, we find that $\gamma = 0.9$ or 0.95 can be a good starting point.

D. Discussion on Our Method's Applicability to Real Human Preference

In the experiments (Section 4), we use human-preference scorers for quantitatively verifying our method's ability to satisfy (human) preferences, which also facilitates reproducibility. Human-preference scorers are also essential for further studies

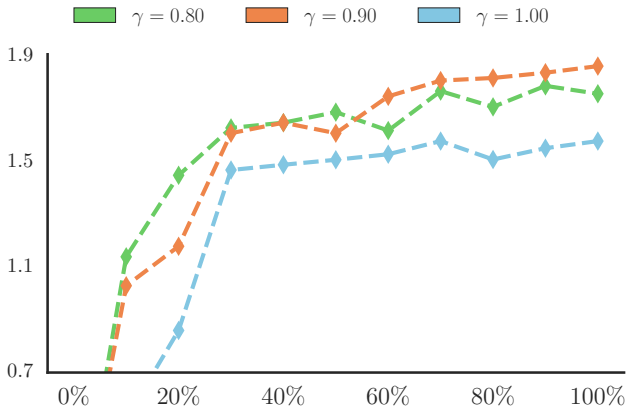


Figure 8: ImageReward over the training process for the single prompt experiment on the color-domain prompt "A green colored rabbit.", with $\gamma \in \{0.8, 0.9, 1.0\}$. As in Fig. 6a, x-axis represents $t\%$ of the training process and all lines start from -0.02 at 0%, the value of "Orig." in Fig. 1a.

of our proposed method in Section 4.3 **(b)** and **(c)**. Apart from the numeric scores, we present image samples and conduct a human evaluation (Section 4.3 **(d)**) to verify our method's ability in generating (human) preferable images.

As presented in Section 2, our method does not make assumptions about the preference source. Thus, a reward function is not an intrinsic requirement of our method. Being agnostic to the preference source, our method is readily applicable to (real) human preferences as well.

Adapting the classical off-policy RLHF paradigm in the literature (*e.g.*, Ziegler et al., 2019; Stiennon et al., 2020; Menick et al., 2022; Bai et al., 2022a), a simple workflow of applying our method to real human preferences iterate on:

1. Generate trajectories from the latest policy, gather human preferences on the corresponding images, and store the quantities required in Section 2;

2. Continue training the T2I by our proposed loss for a chosen number of steps, utilizing the newly collected human data.

Given its similarity with the cited RLHF literature, we believe that this workflow is indeed practical for human-in-the-loop.

E. More Related Works

Dense *v.s.* Sparse Training Guidance for Sequential Generative Models. By its sequential generation nature, T2Is are instances of generative models with *sequential* nature, which further includes, *e.g.*, text generation models, (Devlin et al., 2018; Lewis et al., 2019; Radford et al., 2019) and dialog systems (Chen et al., 2017; Kwan et al., 2022). Similar to T2I's alignment (Section 3), a classical guiding signal for training sequential generative models is the native trajectory-level feedback such as the downstream test metric (*e.g.*, Ryang & Abekawa, 2012; Ranzato et al., 2015; Rennie et al., 2017; Paulus et al., 2017; Shu et al., 2021; Lu et al., 2022; Snell et al., 2022). As discussed in Section 1, ignoring the sequential-generation nature can incur optimization difficulty and training instability due to the sparse reward issue (Guo et al., 2022; Snell et al., 2022). In RL-based methods for training text generation models, in particular, it has become popular to incorporate into the training objective a per-step KL penalty towards the uniform distribution (Guo et al., 2022; Deng et al., 2022), the initial pre-trained model (Ziegler et al., 2019; Ramamurthy et al., 2022), the supervised fine-tuned model (Jaques et al., 2019; Stiennon et al., 2020; Jaques et al., 2020; Ouyang et al., 2022), or some base momentum model (Castricato et al., 2022), to "densify" the sparse reward. Although a per-step KL penalty does help the RL-based training, it can be less task-tailored should one regularizes the generative models towards those generic distributions, especially regarding the ultimate training goal — optimizing the desired trajectory-level feedback. As discussed in Yang et al. (2023) (Appendix F), when combined with the sparse reward issue, such a KL regularization can in fact distract the training of text generation models from improving the received feedback, especially for the initial steps of the generation process, which unfortunately will affect all subsequent generation steps.

In some relatively restricted settings, task-specific *dense* rewards have been explored for training text generation models. With the assumption of abundant expert data for supervised (pre-)training, Shi et al. (2018) use inverse RL (Russell, 1998) to infer a per-step reward; Guo et al. (2018) propose a hierarchical approach; Yang et al. (2018) learn LM discriminators; while

Lin et al. (2017) and Yu et al. (2017) first learn a trajectory-level adversarial reward function similar to a GAN discriminator, before applying the expensive and high-variance Monte-Carlo rollout to simulate per-step rewards. In the code generation domain, Le et al. (2022) use some heuristic values related to the trajectory-level evaluation, without explicitly learning per-step rewards.

Inspired by preference learning in robotics (*e.g.*, Christiano et al., 2017), methods have been recently developed to learn a *dense* per-step reward function whose trajectory-level aggregation aligns with the preference ordering among multiple alternative generations. These methods have been applied to both sufficient-data and low-data regime, in applications of training task-oriented dialog systems (*e.g.*, Ramachandran et al., 2021; Feng et al., 2023) and fine-tuning text-sequence generation models (Yang et al., 2023).

Motivated by this promising direction in prior work and an easier learning problem in RL, in this paper, we continue the research on *dense* training guidance for sequential generative models, by assuming that the trajectory-level preferences are generated by a latent *dense* reward function. Through incorporating the key RL ingredient of temporal discounting factor γ, we break the temporal symmetry in the DPO-style explicit-reward-free alignment loss. Our training objective naturally suits the T2I generation hierarchy by emphasizing the initial steps of the T2I generation process, which benefits all subsequent generation steps and thereby improves both effectiveness and efficiency of training, as shown in our experiments (Section 4).

Characterizing the (Latent) Preference Generation Distribution. Since preference comparisons are typically performed only among the fully-generated trajectories, aligning trajectory generation with preference mostly requires characterizing how preference is originated from per-step rewards, as part of the preference model's assumptions. In the imitation learning literature, preference model is classical chosen to be the Boltzmann distribution over the *undiscounted* sum of per-step rewards (Christiano et al., 2017; Brown et al., 2019; 2020). Several advances have been made on the characterization of the preference model, especially for accommodating the specific nature of concrete tasks. In robotics, Kim et al. (2023) proposes to model the (negative) potentials of the Boltzmann distribution by *learning* a *weighted-sum* to aggregate the per-step rewards over the entire trajectory. Motivated by the simulated robotics benchmark of location/goal reaching, an alternative formulation has been developed that models the potentials of the preference Boltzmann distribution by the optimal advantage function or regret (Knox et al., 2022; 2023; Hejna et al., 2023). Of a special note, though the objective in CPL (Eq. (5) in Hejna et al. (2023)) looks similar to our Eq. (8), **in experiments, CPL actually sets** $\gamma = 1$ (Page 29 Table 6 of Hejna et al. (2023)), making **their actual loss indeed being the "trajectory-level reward" variant** discussed in Section 2.4. Apart from robotic tasks, in text-sequence generation, Yang et al. (2023) take into account the variable-length nature of the tasks, *e.g.*, text summarization, and propose to incorporate inductive bias into modelling the potentials of the preference Boltzmann distribution, through a *task-specific selection* on how the per-step rewards should be aggregated over the trajectory. In this paper, we are among the earliest works to consider the characterization of the preference model in T2I's alignment. By incorporating temporal discounting ($\gamma < 1$) into the preference Boltzmann distribution, we cater for the generation hierarchy of the diffusion and T2I reverse chain (Ho et al., 2020; Wang & Vastola, 2023). Through experiment results and further study (Section 4, especially Section 4.3 **(a)** & **(b)**), we demonstrate that temporal discounting can be useful for effective and efficient T2I preference alignment.

Learning-from-preference in Related Fields. As discussed before, learning-from-preference has been a longstanding problem in robotics/control tasks (Akrour et al., 2011; 2012; Fürnkranz et al., 2012) and has recently been scaled up to train deep-neural-network-based policies (Christiano et al., 2017; Ibarz et al., 2018; Bıyık et al., 2019; Brown et al., 2019; 2020; Lee et al., 2021; Shin et al., 2021; Hejna & Sadigh, 2023a;b). These methods typically start by learning a reward function from data of pairwise comparisons or rankings, before using RL algorithms for policy optimization. Motivated by its success in robotics, learning-from-preference is adopted in the field of natural language generation to improve text summarization (Ziegler et al., 2019; Stiennon et al., 2020) and has become a *de-facto* ingredient in the recent trend of LLMs and conversational agent (*e.g.*, Ouyang et al., 2022; Bai et al., 2022a; Menick et al., 2022; OpenAI, 2023). Apart from the fine-tuning stage, learning-from-preference has also been applied to the pre-training stage, though only use the sparse trajectory-level evaluation (Korbak et al., 2023). To alleviate the modelling and compute complexity of an explicit reward model, following the maximum-entropy principle in control and RL (Ziebart et al., 2008; Ziebart, 2010; Finn et al., 2016), DPO-style objectives (*e.g.*, Rafailov et al., 2023; Tunstall et al., 2023; Azar et al., 2023; Yuan et al., 2023; Zhao et al., 2023; Ethayarajh et al., 2023) directly train the LLMs to align with the preference data, without explicitly learning a deep neural network for the reward function. In this paper, we are among the earliest to study the extension of learning-from-preference into T2I's preference alignment. By taking a dense-reward perspective, we contribute to the DPO-style explicit-reward-free methods by developing a novel objective that emphasizes the initial part of the sequential generation process, which better accommodates the generation hierarchy of diffusion models and T2Is (Ho et al., 2020; Wang & Vastola, 2023). We validate

our perspective through experiments in Section 4.

F. Experiment Details

We note that in mini-batch training of Eq. (8), for the the sampled mini-batch $\mathcal{B} \triangleq \{(\tau_i^1, \tau_i^2)_{\boldsymbol{c}_i}\}_i$, each trajectories in the trajectory tuple (τ_i^1, τ_i^2) corresponds to the same text prompt $\boldsymbol{c}_i$, which makes the preference comparison between trajectories valid. Different trajectory tuples may correspond to different text prompts in the multiple prompt experiments.

We implement our method based on the source code of DPOK (Fan et al., 2023), and inherit as many of their designs and hyperparameter settings as possible, *e.g.*, the specific U-net layers to add LoRA. In the notation of Section 2, the LoRA parameters are our trainable policy parameter θ. To further save GPU memory, the entire training process is conducted under `bfloat16` precision. For training stability, we are motivated by PPO (Schulman et al., 2017) and DPOK to clip all log density ratios $\log \frac{\pi_\theta(a_t \mid s_t)}{\pi_I(a_t \mid s_t)}$ to be within $[-\epsilon, \epsilon]$, since $\log(1 \pm \epsilon) \approx \pm\epsilon$. Without further tuning, we set $\epsilon = 1\mathrm{e}-4$ in single prompt experiments as in DPOK, and $\epsilon = 5e-4$ in multiple prompt experiments.

Below we discuss the hyperparameter settings specific to our single and multiple prompt experiments.

Table 6: Key hyperparameters for T2I (policy) training in the single prompt experiments.

Hyperparameter	Value
M_{tr}	10000
M_{col}	2500
N_{pr}	1000
N_{traj}	5
N_{step}	3
C	10.0
γ	0.9
Batch Size	4
LoRA Rank	4
Optimizer	AdamW
Learning Rate	3e-5
Weight Decay	2e-3
Gradient Norm Clipping	1.0
Learning Rate Scheduler	Constant
Preference Source	ImageReward

Table 7: Key hyperparameters for T2I (policy) training in the multiple prompt experiments.

Hyperparameter	Value
M_{tr}	40000
M_{col}	4000
N_{pr}	2000
N_{traj}	5
N_{step}	1
C	12.5
γ	0.9
Batch Size	32
LoRA Rank	32
Optimizer	AdamW
Learning Rate	2e-5
Weight Decay	1.5e-3
Gradient Norm Clipping	0.05
Learning Rate Scheduler	Constant
Preference Source	HPSv2

F.1. Single Prompt Experiments

Table 6 tabulates the key training hyperparameters, where we use the Adam optimizer with decoupled weight decay (AdamW, Loshchilov & Hutter, 2017).

F.2. Multiple Prompt Experiments

We note that we obtained the HPSv2 train prompts by email correspondence with HPSv2's authors. We produce all results by following the testing principle in the HPSv2 paper and the official GitHub Repository. Table 7 tabulates the key training hyperparameters, where we again use the AdamW optimizer. In the qualitative comparisons (Fig. 4 and Appendix G.2), image samples for the baseline "Dreamlike Photoreal 2.0" are directly from the officially released HPSv2 benchmark images.

F.3. Setups of the Human Evaluation

In our human evaluation (Section 4.3 **(d)**), we generally adopt the principle in prior work (*e.g.*, Wu et al., 2023a; Xu et al., 2023; Wallace et al., 2023a) to evaluate the generated images' fidelity to the text prompt, as well as their overall quality. We use the same set of baseline methods as in Fig. 4, since we view this set as both representative and minimal. In conducting this evaluation, we *randomly sampled* 200 prompts from the HPSv2 test set. Note that though we use the same set of baseline

methods, the sampled text prompts are *not* necessarily the same as those shown in Fig. 4 and Appendix G.2. We asked 20 qualified evaluators for binary comparisons between two images, each from a different model, based on the provided corresponding text prompt. The method names were anonymized. The evaluators were asked to read the text prompt and select which one of the two images is better, in terms of both text fidelity and image quality. To reduce randomness and bias in human judgement, we ensured that all binary comparisons would be evaluated multiple times by the same or a different evaluator. In Table 3, we report the "win rate" of our method, *i.e.*, the percentage of binary comparisons with the stated opponent where the image from our method is preferred. Note that the "win rate" is averaged over all comparisons between the specified two parties. We leave as future work a more comprehensive and larger scale human evaluation for our method.

G. More Generated Images

G.1. More Images from the Single Prompt Experiment

Figure 9: Single prompt experiment: *randomly sampled* generated images for the prompt "A green colored rabbit.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 10: Single prompt experiment: *randomly sampled* generated images for the prompt "Four wolves in the park.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 11: Single prompt experiment: *randomly sampled* generated images for the prompt "A cat and a dog.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 12: Single prompt experiment: *randomly sampled* generated images for the prompt "A dog on the moon.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 13: Single prompt experiment: *randomly sampled* generated images for the prompt "A green colored cat.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 14: Single prompt experiment: *randomly sampled* generated images for the prompt "Four birds in the park.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 15: Single prompt experiment: *randomly sampled* generated images for the prompt "A cat and a cup.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 16: Single prompt experiment: *randomly sampled* generated images for the prompt "A lion on the moon.", from our method and the baselines in Fig. 3. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

G.2. More Images from the Multiple Prompt Experiment

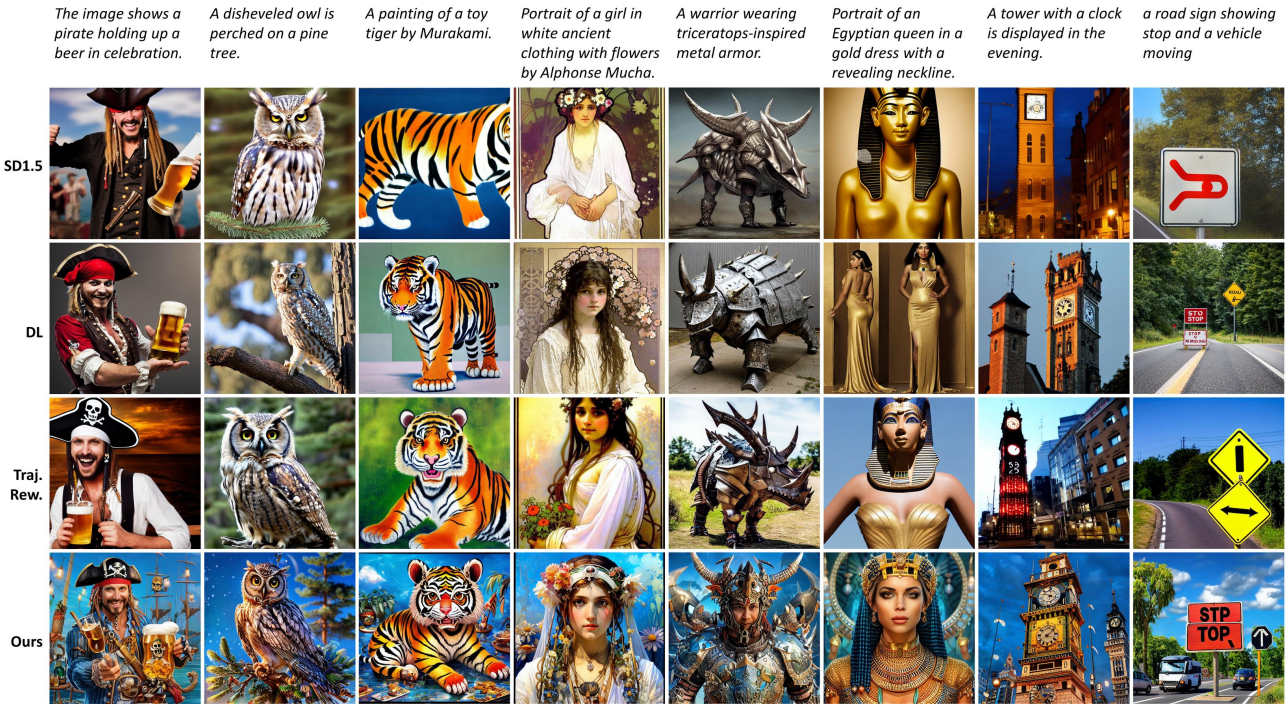


Figure 17: Multiple prompt experiment: generated images from our method and the baselines in Fig. 4 on *randomly sampled* prompts from the HPSv2 test set. "DL" denotes Dreamlike Photoreal 2.0, the best baseline from the HPSv2 paper. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

Figure 18: Multiple prompt experiment: generated images from our method and the baselines in Fig. 4 on *randomly sampled* prompts from the HPSv2 test set. "DL" denotes Dreamlike Photoreal 2.0, the best baseline from the HPSv2 paper. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

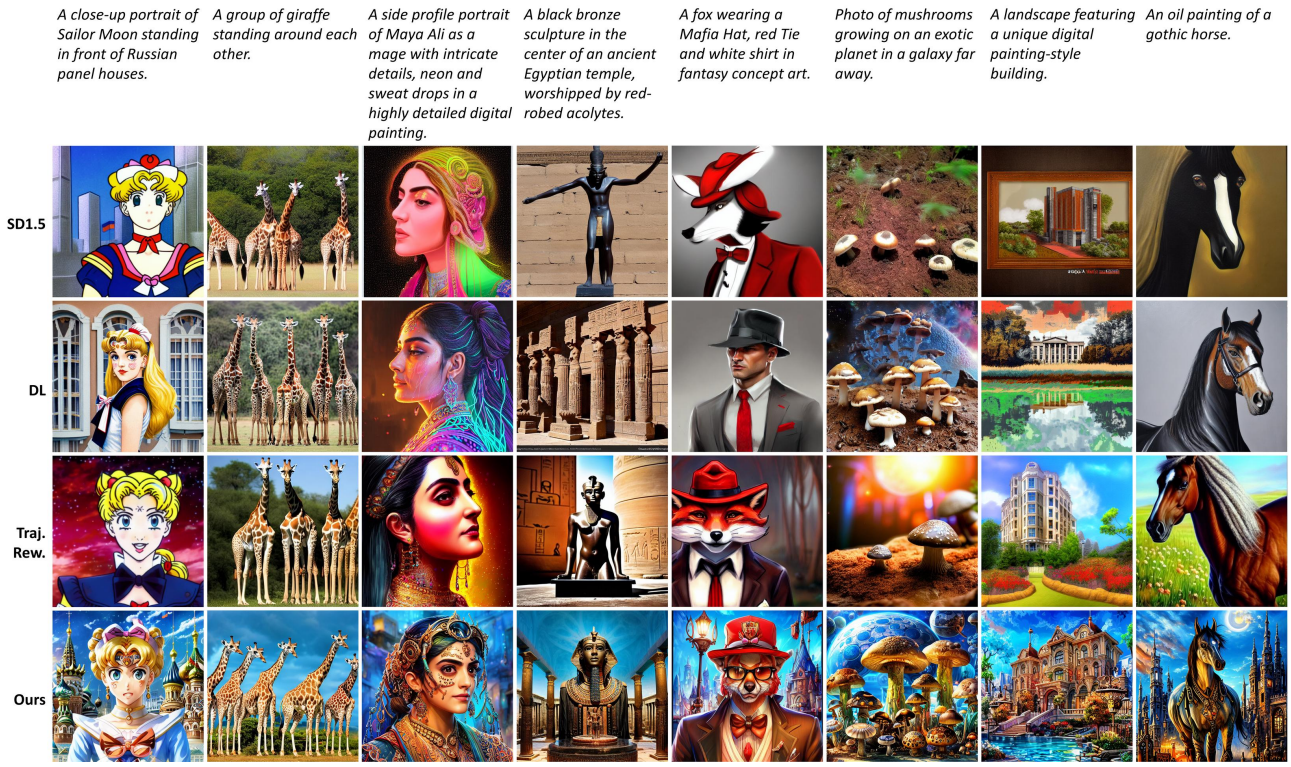


Figure 19: Multiple prompt experiment: generated images from our method and the baselines in Fig. 4 on *randomly sampled* prompts from the HPSv2 test set. "DL" denotes Dreamlike Photoreal 2.0, the best baseline from the HPSv2 paper. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

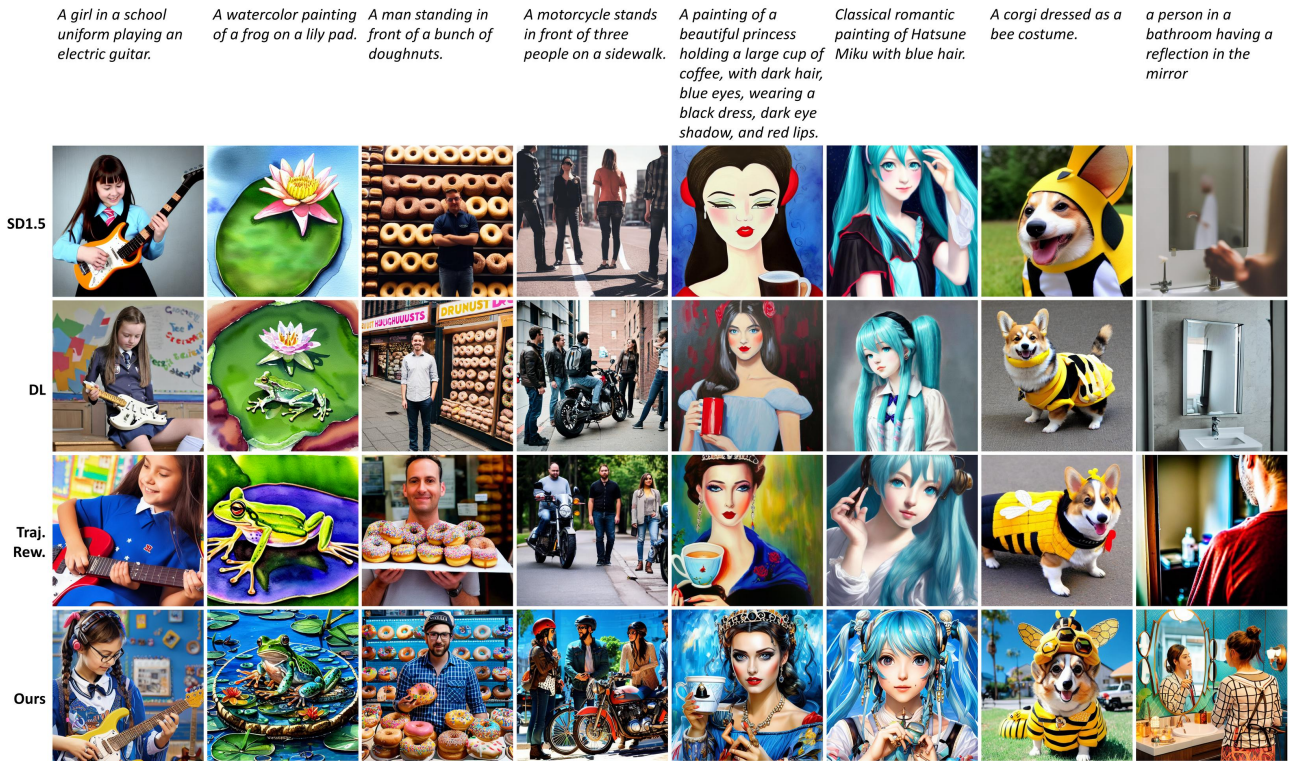


Figure 20: Multiple prompt experiment: generated images from our method and the baselines in Fig. 4 on *randomly sampled* prompts from the HPSv2 test set. "DL" denotes Dreamlike Photoreal 2.0, the best baseline from the HPSv2 paper. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

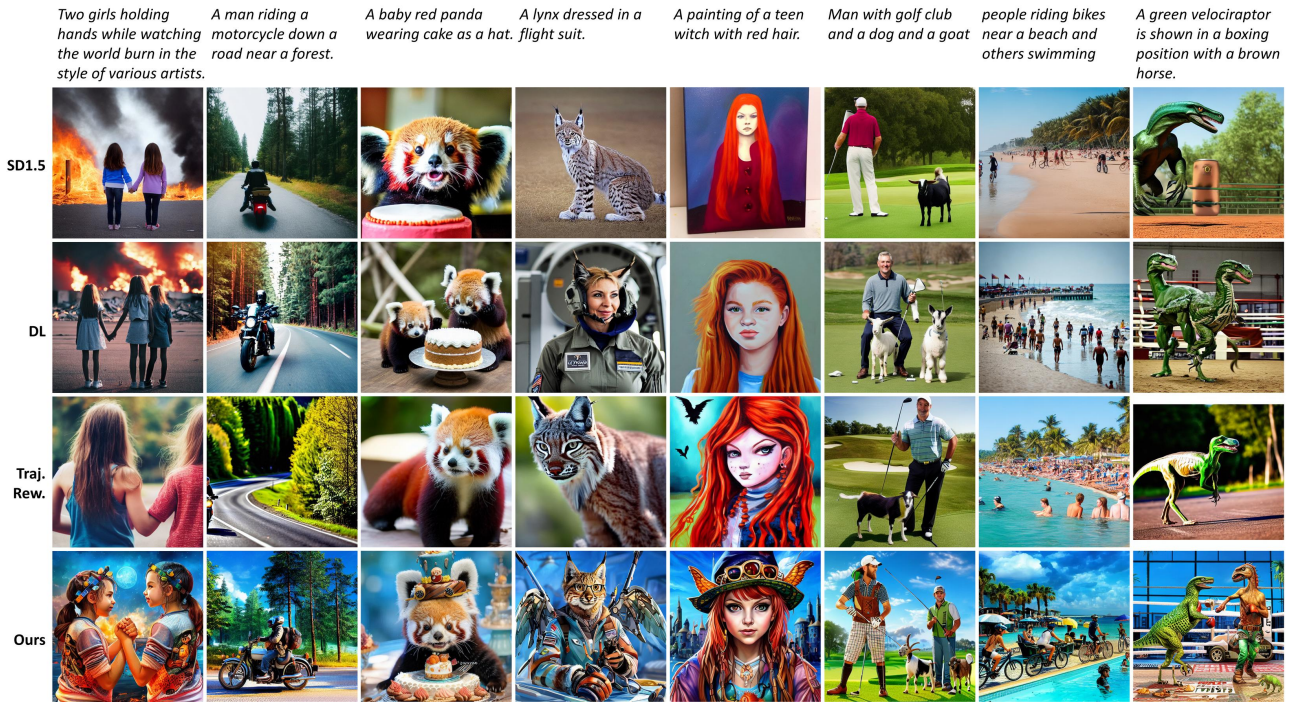


Figure 21: Multiple prompt experiment: generated images from our method and the baselines in Fig. 4 on *randomly sampled* prompts from the HPSv2 test set. "DL" denotes Dreamlike Photoreal 2.0, the best baseline from the HPSv2 paper. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4).

G.3. More Generation Trajectories

Recall that for all generation trajectories, we present the (decoded) $\hat{\boldsymbol{x}}_0$ predicted from the latents at the specified timesteps of the diffusion/T2I reverse chain. A brief discussion on each figure is in its caption.

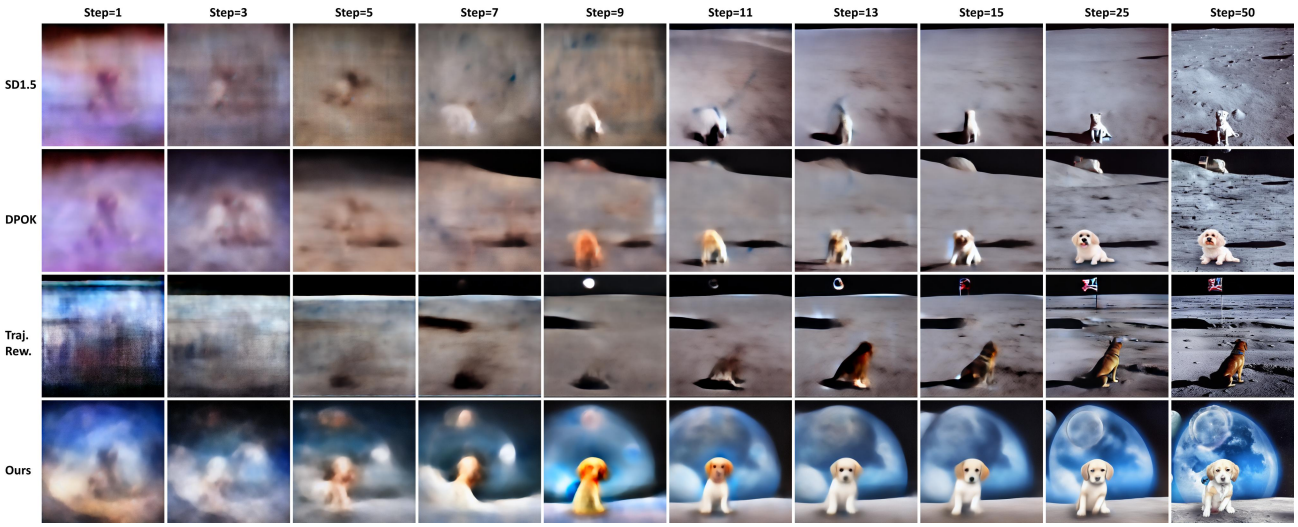


Figure 22: Generation trajectories for the prompt "A dog on the moon.", correspond to the images in Fig. 3 from our method and the baselines. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4). Our method generates the required shape of a *dog* as early as at Step 11, when the shapes in the baselines are mostly unrecognizable. At Step 13, our method is able to give a rather complete generation for the input prompt. Subsequent steps in the reverse chain are then allocated to polish the image details, leading to better final image.

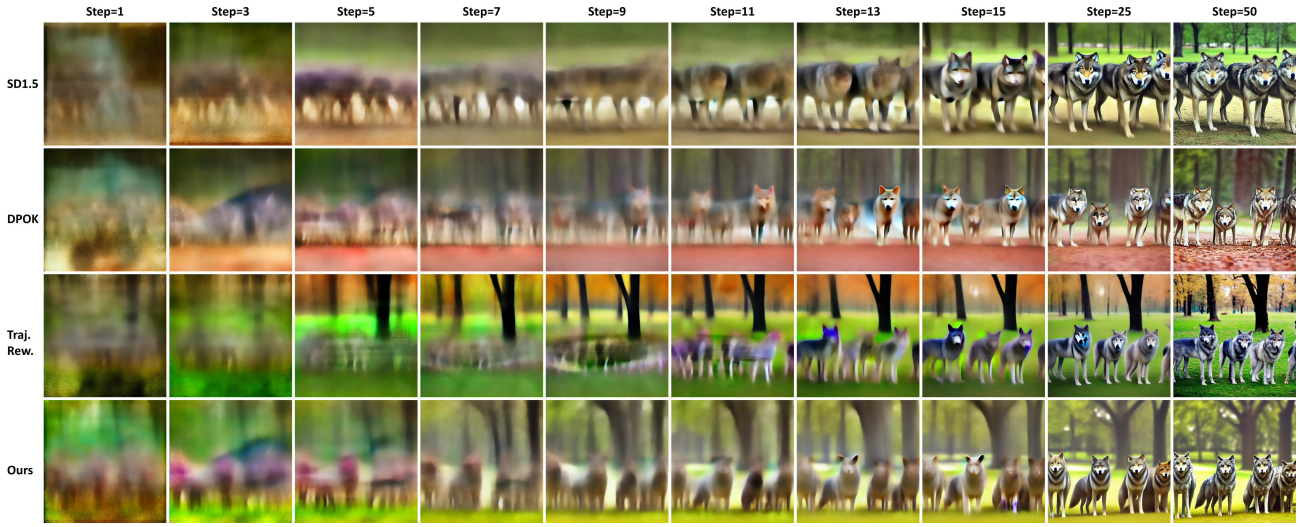


Figure 23: Generation trajectories for the prompt "Four wolves in the park.", correspond to the images in Fig. 3 from our method and the baselines. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4). Our method generates outlines of the requires shapes (four wolves) as early as at Steps 9 and 11, earlier than the baselines especially the "Traj. Rew.".

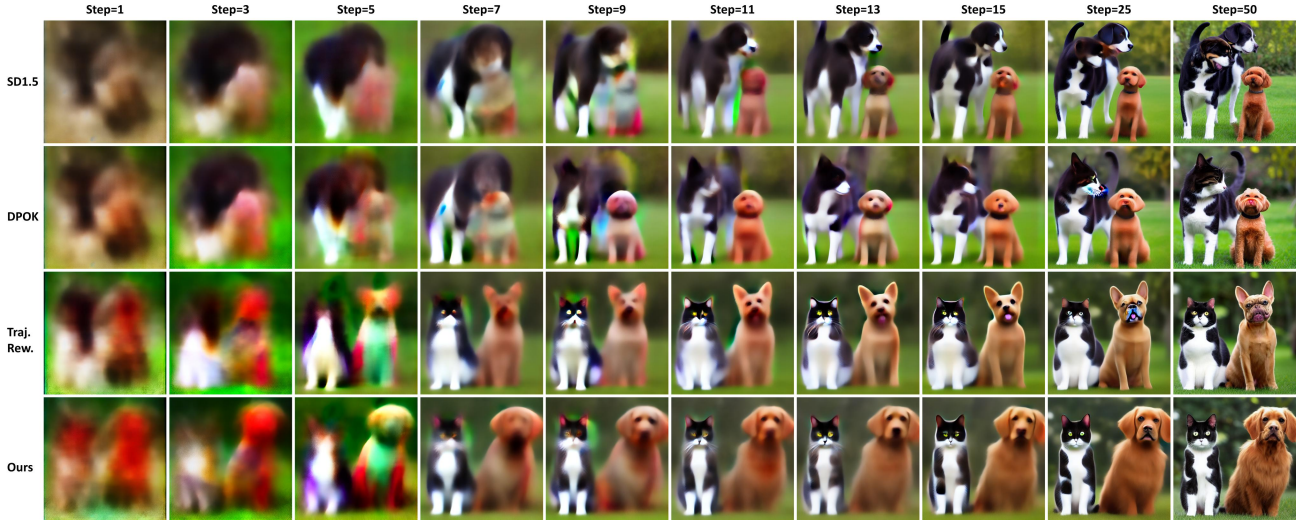


Figure 24: Generation trajectories for the prompt "A cat and a dog.", correspond to the images in Fig. 3 from our method and the baselines. "Traj. Rew." denotes the classical DPO-style objective of assuming trajectory-level reward (Section 2.4). Our method generates the outlines of the desired shapes as fast as at Steps 3 and 5, especially when compared to the baselines DPOK and raw SD1.5. This helps our method in generating a more reasonable and better final image.