Uncertainty-aware Preference Alignment for Diffusion Policies

Runqing Miao[1,2]*, **Sheng Xu**[1]*, **Runyi Zhao**[1], **Wai Kin Victor Chan**[2], **Guiliang Liu**[1]†
[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen,
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University
mrq23@mails.tsinghua.edu.cn, {shengxu1,runyizhao}@link.cuhk.edu.cn
chanw@sz.tsinghua.edu.cn, liuguiliang@cuhk.edu.cn

Abstract

Recent advancements in diffusion policies have demonstrated promising performance in decision-making tasks. To align these policies with human preferences, a common approach is incorporating Preference-based Reinforcement Learning (PbRL) into policy tuning. However, since preference data is practically collected from populations with different backgrounds, a key challenge lies in handling the inherent uncertainties in people's preferences during policy updates. To address this challenge, we propose the Diff-UAPA algorithm, designed for uncertainty-aware preference alignment in diffusion policies. Specifically, Diff-UAPA introduces a novel iterative preference alignment framework in which the diffusion policy adapts incrementally to preferences from different user groups. To accommodate this online learning paradigm, Diff-UAPA employs a maximum posterior objective, which aligns the diffusion policy with regret-based preferences under the guidance of an informative Beta prior. This approach enables direct optimization of the diffusion policy without specifying any reward functions, while effectively mitigating the influence of inconsistent preferences across different user groups. We conduct extensive experiments across both simulated and real-world robotics tasks, and diverse human preference configurations, demonstrating the robustness and reliability of Diff-UAPA in achieving effective preference alignment. The code is available at `https://github.com/mr20010112/Diff_UAPA`.

1 Introduction

Reinforcement Learning (RL) algorithms commonly employ either deterministic or Gaussian policies to tackle sequential decision-making tasks by optimizing cumulative rewards (Wang et al., 2022). Although these RL policies have demonstrated notable success across a wide range of applications (Mnih et al., 2015; Silver et al., 2016; Fang et al., 2019), they may struggle with learning multi-modal policies, which may hinder their ability to generalize effectively and lead to suboptimal performance in complex environments (Zhu et al., 2023). Recently, diffusion models have gained attention due to their strong modeling capabilities (Ho et al., 2020; Song et al., 2020). As a result, more studies have investigated applying diffusion models in RL tasks, particularly in leveraging diffusion models as policies to model complex action distributions (Wang et al., 2023; Chen et al., 2023a; Kang et al., 2023a; Lu et al., 2023; Chi et al., 2023). To learn a diffusion policy that generates desired outputs, recent approaches have leveraged Preference-based Reinforcement Learning (PbRL) (Christiano et al., 2017) techniques, which address a learning-to-rank problem using preference data, enabling alignment with human intentions (Wallace et al., 2024; Dong et al., 2024; Shan et al., 2024).

In practice, preferences are typically gathered from a diverse population, encompassing a wide range of expertise, perspectives, and beliefs. This diversity presents a significant challenge, as preferences

*Equal contributions. †Corresponding author: Guiliang Liu, liuguiliang@cuhk.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

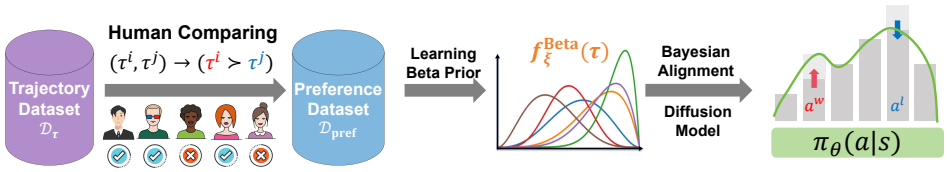

Figure 1: The framework of Diff-UAPA. Given the potentially inconsistent preference dataset ranked by diverse humans, we first learn a Beta prior to capture uncertainties, and then derive a Maximum A Posteriori (MAP) objective to align the diffusion policies.

from different user groups may conflict or evolve over time, introducing great uncertainties during policy updates. To ensure more reliable alignment, this necessitates the development of a policy that could account for the uncertainty arising from potentially inconsistent preferences. However, PbRL approaches are typically based on the Bradley-Terry model (Bradley & Terry, 1952) with maximum likelihood estimation, which lacks sensitivity to the inherent uncertainties from preference datasets.

To address the uncertainties in preference alignment, several methods (Liang et al., 2022; Shin et al., 2023; Xue et al., 2024) have employed techniques such as ensemble models and Bayesian dropout. However, the underlying mechanism by which the estimated ensembles correlate with uncertainty remains largely unexplained. Motivated by a recent work (Xu et al., 2025), which proposes learning a distributional reward model using a Maximum A Posteriori (MAP) objective to address epistemic uncertainty from an offline preference dataset, we explore how to bypass the reward learning and develop an uncertainty-aware algorithm beyond the offline setting for aligning diffusion policies.

In this work, we introduce Uncertainty-aware Preference Alignment for Diffusion Policies (Diff-UAPA), a novel algorithm designed to align diffusion policies with human preferences using an uncertainty-aware objective, as illustrated in Figure 1. Specifically, we introduce an iterative preference alignment framework, in which the diffusion policy progressively adapts to the labels coming from different user groups, each of which may have distinct preferences. To address this challenge, Diff-UAPA involves learning an informative Beta prior that captures the uncertainty arising from diverse human preferences. By interpreting preference alignment as a voting process, we demonstrate that the Beta distribution is sensitive to the uncertainty among trajectories, assigning high confidence to trajectories in which the majority of human raters share a common preference and low confidence to those with divergent preferences. To ensure computational tractability, we parameterize the Beta distribution with neural networks and train the model via variational inference.

Building on the learned Beta prior model, we integrate it into the alignment process with a regret-based preference model, formulating a unified Maximum A Posteriori (MAP) objective. This approach allows for direct optimization of the diffusion policy without the need to learn a reward function, while effectively capturing uncertainty in noisy preferences across diverse user groups.

To assess the empirical performance of Diff-UAPA, we conduct extensive experiments on two simulated robotics environments and one real-world task, comparing it against recent baseline methods. Furthermore, we investigate its effectiveness using diverse human preference data, including synthesized, realistic, and noisy preferences. The results demonstrate the robustness and reliability of Diff-UAPA in handling the uncertainty in preference data.

2 Related Works

2.1 Preference-based Reinforcement Learning

Preference-based Reinforcement Learning (PbRL) is a pivotal approach for aligning agents with human intent, particularly in scenarios where specifying explicit reward functions is challenging (Warnell et al., 2018; Wirth et al., 2017). Previous works generally adopt a two-step procedure, where an explicit reward model is first inferred from human preferences using the Bradley-Terry model (Bradley & Terry, 1952), followed by training an RL agent to optimize the learned reward (Christiano et al., 2017; Ibarz et al., 2018). Building on this framework, several methods (Lee et al., 2021; Park et al., 2022; Liang et al., 2022; Hwang et al., 2023; Choi et al., 2024) have enhanced the learning process, focusing on improving efficiency and capability. While earlier works assume preferences stem from summed Markovian rewards, recent studies (Kim et al., 2023; Verma & Metcalf, 2024) model them using non-Markovian rewards. Another line of work bypasses reward modeling by directly optimizing policies or value functions from human preferences (An et al., 2023; Hejna et al., 2024; Hejna &

Sadigh, 2024). This approach is more straightforward, avoiding the biases and information bottleneck from intermediate reward modeling (Kang et al., 2023b).

2.2 Diffusion Policy for Decision Making

Diffusion models have surpass earlier generative models in both sample quality and training stability, gaining wide attention across domains, including offline RL (Janner et al., 2022; Ajay et al., 2023), online RL (Yang et al., 2023; Chen et al., 2024), and robotics (Sridhar et al., 2024; Chen et al., 2023b). Recent advancements have leveraged diffusion models as RL policies to capture arbitrary action distributions, improving decision-making capabilities (Zhu et al., 2023). Among these works, Diffusion-QL (Wang et al., 2023), first integrated diffusion policies into the Q-learning framework. Following this, SfBC (Chen et al., 2023a) refined policy learning by decoupling behavior learning from action evaluation, while CEP (Lu et al., 2023) extended this framework to enable sampling from broader energy-guided distributions. In preference-based tasks, AlignDiff (Dong et al., 2024) utilized diffusion planners to generate trajectories aligned with human preferences through a two-step procedure, while FKPD (Shan et al., 2024) introduced a one-step framework for direct alignment. However, these methods fail to account for the uncertainties inherent in preferences. How to handle these uncertainties when aligning diffusion policies remains a critical challenge (Casper et al., 2023).

3 Problem Formulation

Preference-based Reinforcement Learning (PbRL). Reinforcement Learning algorithms (Sutton & Barto, 2018) typically consider an episodic Markov Decision Process (MDP), which is formally defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_{\mathcal{R}}, p_{\mathcal{T}}, \gamma, T, \mu_0)$, where: 1) $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, 2) $p_{\mathcal{R}}(r|s,a)$ and $p_{\mathcal{T}}(s'|s,a)$ define the (stochastic) reward and transition functions, 3) $\gamma \in (0,1]$ is the discount factor, 4) μ_0 denotes the initial state distribution and 5) $T \in (0,\infty)$ denotes a non-fixed planning horizon, and the games is reset when the agent reaches a terminating or goal state at a time step T. In many applications, the reward function is not directly available, reducing the episodic MDP to a reward-free MDP $\mathcal{M}_{/r}$. To resolve this challenge, PbRL algorithms Christiano et al. (2017) proposed learning the reward function from human preferences dataset. Specifically, given an unlabeled dataset of trajectory segments $\mathcal{D}_\tau = \{\tau\}$, humans randomly select a pair of trajectories and rank them according to their preferences. By recording these pair-wise comparisons, we create a preference dataset $\mathcal{D}_{\text{pref}} = \{(\tau^w, \tau^l)\}$, where each trajectory segment of length k is defined as $\tau = (s_1, a_1, s_2, a_2, \ldots, s_k, a_k)$, and τ^w is preferred over τ^l. Based on this dataset, recent methods Christiano et al. (2017); Ibarz et al. (2018) commonly infer the rewards by employing the Bradley-Terry model Bradley & Terry (1952) with maximum likelihood estimation (MLE).

Uncertainty Model in Preference Alignment. The Bradley-Terry model (Bradley & Terry, 1952) can effectively model pairwise comparisons, whether by explicitly inferring a reward function (Christiano et al., 2017; Lee et al., 2021; Park et al., 2022) or by directly aligning policies with preferences (Hejna et al., 2024; An et al., 2023). However, this approach fails to account for the inherent uncertainty in human preferences (Newman, 2023; Xu et al., 2025), particularly when these preferences are collected from a diverse population with varying levels of expertise, perspectives, and beliefs. Notably, another line of research focuses on robust preference learning to address noisy labels, employing techniques such as data filtering, label smoothing, and robust loss functions (Cheng et al., 2024; Mandal et al., 2024; Bukharin et al., 2024). However, these approaches often exclude inconsistent data to facilitate training, which risks discarding valuable information if informative data points are mistakenly treated as outliers. More critically, for continuous learning, the policy must adapt dynamically to preferences from different user groups, which often arrive incrementally over time. To resolve these challenges, we study an iterative preference alignment problem:

Definition 3.1 *(Iterative Preference Alignment) Let $\mathcal{D}_\tau = \{\tau\}$ denote the trajectory dataset, and let $\mathcal{D}^n_{pair} = (\tau^i, \tau^j)$ represent the pairwise comparisons dataset constructed at the n^{th} iteration. These comparisons are generated by 1) sampling pairs of trajectories from $\mathcal{D}\tau$ and 2) inviting a group of annotators to label them. The algorithm must progressively align the policy π with the preference dataset $\mathcal{D}^n_{pair}$ at each round $n \in [1, N]$ in an online manner.*

In this setting, different groups of human annotators may provide inconsistent or even conflicting preferences for the same pair of trajectories (Liang et al., 2022; Shin et al., 2023; Xue et al., 2024). The problem solver must dynamically adapt the policy to iteratively updated preference signals while ensuring the learned policy effectively represents general preferences by performing online updates.

In this work, we assume $\mathcal{D}_\tau$ records *offline* trajectories, since interactions with the environments are not always feasible. The primary challenge is to stabilize the policy optimization process and learn a reliable control policy by effectively managing the aleatoric uncertainty inherent in stochastic and potentially inconsistent preference signals on the provided trajectories.

Preference Alignment for Diffusion Policies. Denoising diffusion models (Ho et al., 2020) represent a class of generative models characterized by an iterative diffusion and denoising process, which have gained significant attention in decision-making tasks due to their ability to represent complex distributions (Zhu et al., 2023). This capability is crucial for characterizing the policy function $\pi_\theta(a|s)$, surpassing previous deterministic or Gaussian-based policies (Wang et al., 2023). Diffusion policies are typically formulated as conditional generative models as follows[2]:

$$\pi_\theta(a_t|s_t) = \int \mathcal{N}(a_t^I; \mathbf{0}, \mathbf{I}) \prod_{i=1}^{I} \pi_\theta(a_t^{i-1}|a_t^i, s_t) \mathrm{d}{a_t}^{1:I}, \tag{1}$$

where $\pi_\theta({a_t}^{i-1}|{a_t}^i, s_t)$ is often parameterized as Gaussian with fixed timestep-dependent covariances as $\mathcal{N}(a_t^{i-1}|\mu_\theta(a_t^i, s_t, i), \Sigma^i)$. Although diffusion policies can be trained from expert demonstrations, their performance is often constrained by the size and quality of such dataset. As a result, many previous methods have explored aligning diffusion policies with human feedback (Liu et al., 2024). In this setting, recent research (Wallace et al., 2024) proposed leveraging Direct Preference Optimization (DPO) (Rafailov et al., 2023) to align text-to-image diffusion models with human preferences. Specifically, DPO algorithms directly optimize policies without learning a reward model as follows:

$$L_{\text{DPO}}(\theta) = -\mathbb{E}\left[\log\sigma\left(\lambda\log\frac{\pi_\theta(a^{0,w} \mid s^w)}{\pi_{\text{ref}}(a^{0,w} \mid s^w)} - \lambda\log\frac{\pi_\theta(a^{0,l} \mid s^l)}{\pi_{\text{ref}}(a^{0,l} \mid s^l)}\right)\right], \tag{2}$$

where $\left((s^w, a^{0,w}), (s^l, a^{0,l})\right) \sim \mathcal{D}_{\text{pref}}$ are state-action samples and π_{ref} denotes the reference policy. In this study, we investigate the alignment of diffusion policies with human preferences within the context of reinforcement learning, addressing the challenge of iterative preference alignment.

4 Uncertainty-Aware Preference Alignment for Diffusion Policies

In this section, we outline our approach for aligning a diffusion policy with human preferences while effectively accounting for uncertainty. Specifically, we present: 1) a Maximum Likelihood Estimation (MLE) objective for diffusion policy alignment in the context of RL, based on maximum entropy framework and direct preference optimization (Section 4.1), 2) a Maximum A Posteriori (MAP) objective that incorporates a Beta prior model for capturing the underlying uncertainties (Section 4.2), and 3) the training procedure for the Beta prior model (Section 4.3).

4.1 Maximum Likelihood Diffusion Policy Alignment

MaxEnt Alignment under Regret Preference. Following previous works on preference alignment (Hejna et al., 2024; Rafailov et al., 2024), we adopt the Maximum Entropy (MaxEnt) RL framework (Ziebart, 2010). The objective is to learn a policy π_θ that maximizes both the cumulative discounted rewards and causal entropy, while regularizing the KL-divergence from a reference policy:

$$\max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t(r(s_t, a_t) - \alpha\log\frac{\pi(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)})\right], \tag{3}$$

Here, α determines the weight of entropy. Upon learning an optimal policy π^*, we can compute the corresponding optimal state-value function $V^*(s_t)$, the optimal state-action value function $Q^*(s_t, a_t)$, and the optimal advantage function $A^*(s_t, a_t) \triangleq Q^*(s_t, a_t) - V^*(s_t)$, which measures the relative benefit of taking a specific action in a given state compared to the average expected value of the state. More importantly, in the MaxEnt RL setting, the optimal advantage function is proportional to the log-likelihood of the optimal and reference policy (Haarnoja et al., 2017; Hejna et al., 2024):

$$A^*(s_t, a_t) = \alpha\log\frac{\pi^*(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}. \tag{4}$$

To stabilize the process of preference alignment, we follow Knox et al. (2022) and base the preference alignment on discounted regrets, defined as $-\sum\gamma^t\left(V(s_t) - Q(s_t, a_t)\right)$. In this framework, a

[2]In this work, we use superscripts ($i \in \{0, 1, \ldots, I\}$ to denote diffusion timesteps and subscripts ($t \in \{0, 1, \ldots, T\}$) to denote trajectory timesteps.

trajectory segment is preferred if it incurs lower regret compared to the intended optimal policy, so that the preference between trajectory segments (τ^w, τ^l) can be modeled as:

$$P_{A*}(\tau^w \succ \tau^l) = \frac{\exp \sum_{t=0}^{T} \gamma^t A^*(s_t^w, a_t^w)}{\exp \sum_{t=0}^{T} \gamma^t A^*(s_t^w, a_t^w) + \exp \sum_{t=0}^{T} \gamma^t A^*(s_t^l, a_t^l)}. \tag{5}$$

By substituting Equation (4) into Equation (5), the advantage function A^* can be replaced by the optimal policy π^* under the MaxEnt framework. The learned policy π_θ can then be optimized through maximum the likelihood of generating preferences as follows (Hejna et al., 2024):

$$\mathcal{L}_{\text{CPL}}^{(\tau^w, \tau^l)}(\theta) = -\log \sigma \Big(\alpha \cdot \big(\sum_{t=0}^{T} \gamma^t \log \frac{\pi_\theta(a_t^w | s_t^w)}{\pi_{\text{ref}}(a_t^w | s_t^w)} - \sum_{t=0}^{T} \gamma^t \log \frac{\pi_\theta(a_t^l | s_t^l)}{\pi_{\text{ref}}(a_t^l | s_t^l)} \big) \Big), \tag{6}$$

Diffusion Policy Alignment. To adapt the previous model to aligning the diffusion policy $\pi_\theta(a_t|s_t)$ as defined in Eq. (1), a primary difficulty is due to the intractability of diffusion policy $\pi_\theta(a_t|s_t) = \int \pi_\theta(a_t^{0:I}|s_t) \mathrm{d}a_t^{1:I}$, as it requires marginalizing over all possible diffusion paths $(a_t^1, a_t^2, \ldots, a_t^I)$ that lead to a_t^0. To address it, we propose modeling the chain reward function (Wallace et al., 2024):

$$r(s_t, a_t^0) = \mathbb{E}_{\pi_\theta(a_t^{1:I} | a_t^0, s_t)}[r(s_t, a_t^{0:I})]. \tag{7}$$

The optimal chain advantage function can be defined by marginalizing intermediate diffusion path:

$$A^*(s_t, a_t^0) = \mathbb{E}_{\pi_\theta^*(a_t^{1:I} | a_t^0, s_t)} \left[A^*(s_t, a_t^{0:I}) \right] = \mathbb{E}_{\pi_\theta^*(a_t^{1:I} | a_t^0, s_t)} \left[\alpha \log \frac{\pi_\theta^*(a_t^{0:I}|s_t))}{\pi_{\text{ref}}(a_t^{0:I}|s_t))} \right]. \tag{8}$$

In principle, we can interpret the latent diffusion actions as a unified chain action $\overline{a_t} = a_t^{0:I}$, despite the final output being determined by a_t^0. This perspective allows us to reformulate Equation (3) in terms of the diffusion policy:

$$\max_{\pi_\theta} \mathbb{E}_{\pi_\theta(\overline{a_t}|s_t)} \left[\sum_{t=0}^{T} \gamma^t (r(s_t, \overline{a_t}) - \alpha \log \frac{\pi_\theta(\overline{a_t}|s_t))}{\pi_{\text{ref}}(\overline{a_t}|s_t))} \right]. \tag{9}$$

This objective is defined over the entire diffusion path $\overline{a_t}$, which aims to maximize the cumulative rewards and the entropy within a trajectory across the reverse process. By paralleling from Equation (3) to Equation (6), the objective in (9) can be directly optimized with respect to the diffusion policy $\pi_\theta(\overline{a_t}|s_t)$ by maximizing the following likelihood:

$$\begin{aligned} \mathcal{L}_{\text{1,MLE}}^{(\tau^w, \tau^l)}(\theta) = -\log \sigma \Big(\alpha \cdot \Big(& \sum_{t=0}^{T} \mathbb{E}_{\pi_\theta(a_t^{1:I,w} | s_t^w, a_t^{0,w})} \left[\gamma^t \log \frac{\pi_\theta(\overline{a_t^w}|s_t^w)}{\pi_{\text{ref}}(\overline{a_t^w}|s_t^w)} \right] \\ & - \sum_{t=0}^{T} \mathbb{E}_{\pi_\theta(a_t^{1:I,l} | s_t^l, a_t^{0,l})} \left[\gamma^t \log \frac{\pi_\theta(\overline{a_t^l}|s_t^l)}{\pi_{\text{ref}}(\overline{a_t^l}|s_t^l)} \right] \Big) \Big), \end{aligned} \tag{10}$$

where σ is the sigmoid function. However, major challenges in optimizing this objective are: 1) *inefficiency* from sequential computations over timesteps, and 2) *intractability* due to evaluating the joint distribution. Following Wallace et al. (2024), we apply Jensen's inequality and the convexity of $-\log \sigma$ to move the expectation outside, boosting efficiency. We also approximate the reverse process $\pi_\theta(a_t^{1:I}|s_t)$ with the forward process $q(a_t^{1:I}|s_t)$ to enhance tractability.

With some algebra, we derive the following loss function:

$$\begin{aligned} \mathcal{L}_{\text{1,MLE}}^{(\tau^w, \tau^l)}(\theta) \leqslant & -\mathbb{E}_{\substack{a_t^{i,w} \sim q(a_t^{i,w} | a_t^{0,w}, s_t^w), \\ a_t^{i,l} \sim q(a_t^{i,l} | a_t^{0,l}, s_t^l)}} \Big[\log \sigma \Big(-\alpha I \cdot \\ & \Big(\sum_{t=0}^{T} \gamma^t (\|\epsilon^w - \epsilon_\theta(a_t^{i,w}, s_t^w, i)\|_2^2 - \|\epsilon^w - \epsilon_{\text{ref}}(a_t^{n,w}, s_t^w, i)\|_2^2) \\ & - \sum_{t=0}^{T} \gamma^t (\|\epsilon^l - \epsilon_\theta(a_t^{i,l}, s_t^l, i)\|_2^2 - \|\epsilon^l - \epsilon_{\text{ref}}(a_t^{i,l}, s_t^l, i)\|^2) \Big) \Big) \Big] = \mathcal{L}_{\text{2,MLE}}^{(\tau^w, \tau^l)}(\theta), \end{aligned} \tag{11}$$

where 1) $i \sim \mathcal{U}(0, I)$ is the diffusion timestep, 2) $a^{i,w/l} \sim q(a^{i,w/l}|a^{0,w/l}, s^{w/l})$ denotes the action $a^{0,w/l}$ corrupted with noise $\epsilon^{w/l}$ after i diffusion steps, as defined in (Ho et al., 2020). In this study, we explore addressing the iterative preference alignment problem by aligning human preferences with a diffusion policy model. The detailed deviation is shown in Appendix A.

4.2 Bayesian Alignment with Informative Beta Prior

The regret preference model (Eq. (5)) represents the likelihood of generating human preferences based on the advantage function. The corresponding maximum likelihood objective implicitly assumes a uniform prior over $\sum_{t=0}^{T} \gamma^t A^*(s_t, a_t)$, which does not account for the uncertainty within the preference dataset (Newman, 2023; Xu et al., 2025). We next derive a more informative prior.

Since human feedback is provided at the trajectory level rather than for individual state-action pairs, the strength of a trajectory can be defined by its trajectory-level advantage, computed as the discounted cumulative advantage under the diffusion policy π_θ:

$$A^{\pi_\theta}(\tau) = \sum_{t=0}^{T} \gamma^t A^{\pi_\theta}(s_t, a_t) = \sum_{t=0}^{T} \gamma^t \mathbb{E}_{\pi_\theta(a_t^{1:I}|a_t^0, s_t)} \left[A^{\pi_\theta}(s_t, \overline{a_t}) \right]. \tag{12}$$

The average strength of the trajectories under policy π_θ is then defined as:

$$\bar{A}^{\pi_\theta} = \mathbb{E}_{\tau \sim \mathcal{D}_\tau} A_\theta(\tau) = \frac{1}{|\mathcal{D}_{\text{pref}}|} \sum_{\tau \in \mathcal{D}_{\text{pref}}} A^{\pi_\theta}(\tau). \tag{13}$$

Therefore, we model the binary outcome of whether a trajectory with strength $A^{\pi_\theta}(\tau)$ wins against the average candidate as a Bernoulli variable with success probability $\phi(\tau) = \sigma(A^{\pi_\theta}(\tau) - \bar{A}^{\pi_\theta}) \in (0, 1)$.

By applying the chain rule, the prior on the advantage function can be defined as:

$$p_0(A^{\pi_\theta}(\tau)) = p_0(\phi(\tau)) \frac{\mathrm{d}\phi(\tau)}{\mathrm{d}A^{\pi_\theta}(\tau)} = p_0(\phi(\tau)) \sigma'(A^{\pi_\theta}(\tau) - \bar{A}^{\pi_\theta})(1 - \frac{1}{|\mathcal{D}_{\text{pref}}|}). \tag{14}$$

This prior reflects our initial belief about the strength of different trajectories within the dataset. Motivated by Xu et al. (2025), we use the Beta distribution $p_0(\phi(\tau)) = \text{Beta}(\phi(\tau); \alpha, \beta)$ as the informative prior, which serves as the probability density function of $\phi(\tau)$. The main benefits of the Beta distribution are: 1) it is the conjugate prior for the Bernoulli distribution $\phi(\tau)$, and 2) the parameters α and β can intuitively represent the counts of *preferred* and *unpreferred* human feedback. By reformulating Eq. (14), we present the following proposition:

Proposition 4.1 *Let the informative prior $p_0(\phi(\tau))$ be a Beta distribution $Beta(\phi(\tau); \alpha, \beta)$. This prior can effectively capture the uncertainty arising from the iterative preference alignment process (Definition 3.1). Consequently, the prior on the strength of a trajectory is proportional to Beta$((\phi(\tau); \alpha + 1, \beta + 1))$, i.e., $p_0(A^{\pi_\theta}(\tau)) \propto$ Beta$(\phi(\tau); \alpha + 1, \beta + 1)$.*

The proof is shown in Appendix C. The corresponding prior loss can then be derived in a manner similar to the derivation of the maximum likelihood loss (Eq. 10):

$$\begin{aligned}
\mathcal{L}^{\tau}_{1,\text{prior}}(\theta) &= -\log \text{Beta}(\phi(\tau); \alpha + 1, \beta + 1) \\
&\leqslant -\mathbb{E}\Bigg[\log \text{Beta}\Bigg(\sigma\Bigg(-\alpha I \cdot \Big(\sum_{t=0}^{T} \gamma^t (\left\| \epsilon - \epsilon_\theta(a_t^i, s_t, i) \right\|_2^2 - \left\| \epsilon - \epsilon_{\text{ref}}(a_t^i, s_t, i) \right\|_2^2) \\
&\qquad - \sum_{\tau \in \mathcal{D}_{\text{pref}}, t=0}^{T} \frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|} (\left\| \epsilon - \epsilon_\theta(a_t^i, s_t, i) \right\|^2 - \left\| \epsilon - \epsilon_{\text{ref}}(a_t^i, s_t, i) \right\|^2) \Big) \Bigg); \alpha + 1, \beta + 1 \Bigg) \Bigg] \\
&= \mathcal{L}^{\tau}_{2,\text{prior}}(\pi_\theta)
\end{aligned} \tag{15}$$

Equation (15) can be interpreted as guiding the diffusion policy to align its estimate of $\phi(\tau)$ with the corresponding prior Beta distribution. According to Bayes' rule, the MAP estimate satisfies $P_{\text{MAP}}(A(\tau)) \propto p_0(A(\tau)) \cdot P_{\text{MLE}}(A(\tau))$. By incorporating the prior into the MLE objective and maximizing the log form of the posterior, we can derive the Diff-UAPA loss:

$$\mathcal{L}_{\text{Diff-UAPA}}(\theta) = \mathbb{E}_{(\tau^w, \tau^l) \sim \mathcal{D}_{\text{pref}}} \Big[\mathcal{L}^{(\tau^w, \tau^l)}_{2,\text{MLE}}(\pi_\theta) + \mathcal{L}^{\tau^w}_{2,\text{prior}}(\pi_\theta) + \mathcal{L}^{\tau^l}_{2,\text{prior}}(\pi_\theta) \Big]. \tag{16}$$

Maximizing the posterior probability incorporates prior knowledge and regularizes advantage values, preventing divergence. We introduce how to estimate the Beta prior in the following section.

4.3 Training the Beta Prior Model

To learn the Beta prior $p_0(\phi(\tau)|\mathcal{D}_{\text{pref}}) = \text{Beta}(\phi(\tau); \alpha, \beta)$ in continuous spaces, following Xu et al. (2025), we propose using a variational inference approach to approximate it by estimating the approximate posterior $q_\xi(\phi(\tau)|\mathcal{D}_{\text{pref}})$, i.e., $p_0(\phi(\tau)|\mathcal{D}_{\text{pref}}) \simeq q_\xi(\phi(\tau)|\mathcal{D}_{\text{pref}})$, where ξ is the model

parameters. The objective is to maximize the Evidence Lower Bound between the prior and posterior. This leads to the following trajectory-wise objective (Xu et al., 2025):

$$\max_{\xi} \mathbb{E}_{\tau}\Big[\mathbb{E}_{q_\xi,(\tau^w,\tau^l)\in\mathcal{D}_{\text{pref}}}[\log \phi(\tau^w)] - \mathbb{E}_{q_\xi,(\tau^w,\tau^l)\in\mathcal{D}_{\text{pref}}}[\log \phi(\tau^l)] - D_{\text{KL}}[q_\xi(\phi(\tau)|\tau) \parallel p(\phi(\tau))]\Big], \tag{17}$$

where 1) $q_\xi(\phi(\tau)|\tau) = \text{Beta}(\alpha_\tau, \beta_\tau)$, where $[\alpha_\tau, \beta_\tau] = f_\xi^{\text{Beta}}(\tau)$ and f_ξ^{Beta} denotes a neural network, 2) $p(\phi(\tau)) = \text{Beta}(\alpha_0, \beta_0)$, with α_0, β_0 specifying our prior belief (we set $\alpha_0 = \beta_0 = 1$ in this work), and 3) $\phi(\tau)$ represents the Bernoulli probability that τ^w is ranked higher than τ^l. The first two terms aim to optimize the parameter ξ to align with the preference dataset, while the final KL-divergence term ensures the posterior distribution does not deviate too far from the prior belief, which can be optimized using the Dirichlet VAE approach (Joo et al., 2020).

In this work, we implement $f_\xi^{\text{Beta}}(\tau)$ using a transformer-based neural network (Vaswani, 2017), where the trajectory τ is fed as input and $[\alpha_\tau, \beta_\tau]$ is produced as the output to form the Beta prior distribution. The complete Diff-UAPA algorithm is shown in Algorithm 1.

Algorithm 1 Uncertainty-aware Preference Alignment for Diffusion Policies (Diff-UAPA)

1: **Input:** Preference dataset $\mathcal{D}_{\text{pref}}$
2: Initialize Beta prior model $f_\xi^{\text{Beta}}(\tau)$, reference policy $\pi_{\text{ref}}(a|s)$, and diffusion policy $\pi_\theta(a|s)$.
3: Learn π_{ref} via behavior cloning on $\mathcal{D}_{\text{pref}}$.
4: Update Beta prior f_ξ^{Beta} via Eq. (17) until convergence.
5: Update π_θ by minimizing Eq. (16) until convergence.

5 Empirical Evaluation

In this section, we perform empirical evaluations on five robot manipulation tasks across two environments (Sec. 5.1), locomotion tasks with real human preferences (Sec. 5.2), and a real-world pick-and-place task (Sec. 5.3). We further conduct more comprehensive analyses, including ablation studies, evaluations under various noise types, and noise sensitivity tests (Sec.5.4).

Experiment Settings. Our experiments consist of four rounds of iterative updates, each with a fixed number of training episodes. To account for potential inconsistencies in human preferences, we introduce a reverse rate. Specifically, in each round, we randomly select 20% of trajectory pairs and apply a 50% reversal rate by flipping the labels. Each experiment is repeated using three random seeds. More experimental details can be found in Appendix D.1.

Comparison Methods. We utilize two baseline policies: the Gaussian-based policy from Behavior Transformer (**BET**) (Shafiullah et al., 2022) and the Diffusion Policy (**Diff**) (Chi et al., 2023). Building on BET, we propose the following comparison methods: 1) BET-Direct Preference Optimization (**BET-DPO**) and 2) BET-Contrastive Preference Learning (**BET-CPL**), which leverage direct preference optimization (Rafailov et al., 2023) and contrastive preference learning (Hejna et al., 2024) to align the BET model and 3) **UA-PbRL**(Xu et al., 2025), which learns a distributional policy for epistemic uncertainty awareness.. For diffusion-based policies, we introduce: 4) Diffusion Policy-CPL (**Diff-CPL**) that uses MLE for aligning the diffusion policy (Obj. 11), and 5) **FKPD** Shan et al. (2024) that performs forward KL regularized preference optimization. For our Diff-UAPA algorithm, we explore two strategies for updating the Beta prior model: 6) **Diff-UAPA-C** that trains the Beta model using full preference data across the iterations without updates and 7) **Diff-UAPA-I** that incrementally updates the Beta model on the current preference data through the iterative process.

5.1 Model Performance in Robot Manipulation Tasks

We evaluate the model's performance across four tasks from Robomimic (Mandlekar et al., 2021) and the Franka Kitchen task introduced in (Gupta et al., 2019), both of which use state-based observations. The Robomimic tasks cover diverse manipulation skills, while the Kitchen task involves long-horizon, multi-step interactions with seven objects, aiming to complete as many tasks as possible in any order. For each task, the reference policy π_{ref} is trained to reach approximately 40% success, then rolled out to collect 560 trajectories used to build the preference dataset based on their rewards. Please check Appendix D.2 for environmental details and Appendix D.4 for preference dataset construction.

Table 1 presents the evaluation results. The results indicate that both variants of Diff-UAPA consistently outperform other methods across different tasks. This is primarily due to their use of a Beta prior, which effectively captures the uncertainty arising from inconsistent preferences, thereby enhancing the diffusion policy training process. Moreover, the performance gap between Diff-UAPA-C and Diff-UAPA-I is relatively small, suggesting that the Beta prior can be trained effectively in

Table 1: Success rates (in percentage) across tasks, with each value presented as the mean $\pm$ std over 3 training seeds and 560 evaluation episodes. The best results for each task are highlighted in bold. For the Kitchen task, px indicates the frequency of interaction with x or more objects.

	Robomimic				Kitchen			
	Lift	Can	Square	Transport	p1	p2	p3	p4
BET	43.6 ± 3.8	48.8 ± 3.1	55.1 ± 2.0	43.1 ± 1.9	96.4 ± 1.2	96.2 ± 1.0	76.6 ± 1.3	44.6 ± 2.0
BET-CPL	49.2 ± 4.4	42.1 ± 1.1	57.6 ± 2.3	45.2 ± 4.8	97.0 ± 1.0	96.4 ± 0.5	88.4 ± 2.3	62.6 ± 2.0
BET-DPO	43.7 ± 3.3	47.0 ± 1.0	42.7 ± 3.6	41.2 ± 2.4	85.5 ± 8.5	84.8 ± 8.7	80.9 ± 9.4	57.4 ± 6.6
UA-PbRL	50.3 ± 2.2	54.4 ± 2.3	55.3 ± 2.7	53.6 ± 5.7	$\mathbf{100.0 \pm 0.0}$	98.7 ± 1.3	92.2 ± 3.7	62.3 ± 4.2
Diff	45.1 ± 3.0	47.9 ± 2.3	52.8 ± 2.9	56.4 ± 3.2	99.2 ± 0.8	98.4 ± 1.1	91.8 ± 0.8	59.0 ± 1.1
Diff-CPL	48.6 ± 2.2	45.9 ± 2.8	55.2 ± 5.7	58.1 ± 6.2	$\mathbf{100.0 \pm 0.0}$	99.6 ± 0.2	94.2 ± 0.2	63.5 ± 0.8
FKPD	51.2 ± 0.7	58.5 ± 2.5	64.4 ± 2.7	52.3 ± 3.5	99.8 ± 0.3	98.3 ± 1.4	89.5 ± 2.9	64.1 ± 3.2
Diff-UAPA-C	$\mathbf{56.1 \pm 0.9}$	$\mathbf{61.3 \pm 2.2}$	$\mathbf{68.1 \pm 0.6}$	$\mathbf{64.0 \pm 4.0}$	$\mathbf{100.0 \pm 0.0}$	99.7 ± 0.2	95.4 ± 0.6	70.9 ± 2.5
Diff-UAPA-I	54.3 ± 1.1	59.9 ± 1.7	66.2 ± 1.3	61.5 ± 2.6	99.9 ± 0.1	$\mathbf{99.8 \pm 0.2}$	$\mathbf{95.7 \pm 1.9}$	$\mathbf{71.7 \pm 4.6}$

both approaches, depending on the specific practice. Notably, for the long-horizon Kitchen task, Diff-UAPA-I, which trains the Beta model incrementally, slightly outperforms Diff-UAPA-C, which pre-trains the Beta model using the complete dataset. This difference can be attributed to the fact that incremental training allows the model to adapt more dynamically to the changing preferences and environmental conditions over time, whereas pre-training may not fully capture such variability. We also provide the visualization results in Figure 3 in Appendix D.6 and Supplementary Materials.

5.2 Model Performance in Locomotion Tasks with Real Human Preferences

The primary goal of PbRL is to align policies with *human* preferences. In this section, we evaluate on real human preferences from Uni-RLHF (Yuan et al., 2024) in the HalfCheetah and Hopper tasks from D4RL (Fu et al., 2020), using *medium-expert* and *medium-replay* datasets to ensure diverse trajectory coverage. Please check Appendix D.2 for more environmental details.

Table 2: Episodic rewards of all methods in the HalfCheetah and Hopper environments with real human preferences. 'M' denotes Medium and 'R' denotes Replay.

	BET	BET-CPL	BET-DPO	Diff	Diff-CPL	FKPD	Diff-UAPA-C	Diff-UAPA-I
HalfCheetah-M-E	2577 ± 198	2976 ± 66	2948 ± 37	2838 ± 325	3121 ± 148	3060 ± 201	$\mathbf{3399 \pm 72}$	3297 ± 101
HalfCheetah-M-R	1580 ± 85	1818 ± 201	1659 ± 198	1691 ± 128	1862 ± 107	1866 ± 203	$\mathbf{2021 \pm 85}$	1949 ± 53
Hopper-M-E	1161 ± 90	1226 ± 85	1129 ± 79	1296 ± 137	1313 ± 103	1370 ± 120	$\mathbf{1591 \pm 51}$	1499 ± 70
Hopper-M-R	702 ± 66	769 ± 34	712 ± 51	780 ± 31	796 ± 20	874 ± 39	$\mathbf{933 \pm 21}$	865 ± 39

The results are presented in Table 2. We find that baseline methods can hardly outperform Diff-UAPA. This occurs because the inconsistent labels increase uncertainty and hinder the policy's ability to identify and imitate better trajectories. Diff-UAPA addresses this by using a prior model to capture uncertainty, enabling more reliable trajectory evaluation and improved performance. We also find that diffusion-based policies outperform Gaussian ones, due to their stronger modeling capacity.

5.3 Real-world Pick-and-Place Task

To assess the effectiveness of the proposed method in sim-to-real transfer and real-world deployment, we conducted a real-world *pick-and-place* experiment using a Rokae SR3 robotic arm. The task involved picking up a banana and placing it in a designated area. Over 3000 simulated trajectories were generated via motion planning with visual observations and proprioceptive inputs, encompassing both expert and sub-optimal executions. Preference labels were assigned to favor expert trajectories.

We conducted real-world evaluations, where our method achieved a 60% success rate, surpassing the 40% achieved by the Diff-CPL baseline. Diff-CPL suffers from the noise that prefers sub-optimal behaviors like reaching the wrong position. As shown in Figure 2, the red circle marks an incorrect pick pose by Diff-CPL, while the green circle highlights the accurate pick-and-place behavior generated by our method. These results demonstrate the robustness and effectiveness of our approach in sim-to-real transfer. We also provide the full process visualization in Figure 4, and videos in the Supplemental Materials.

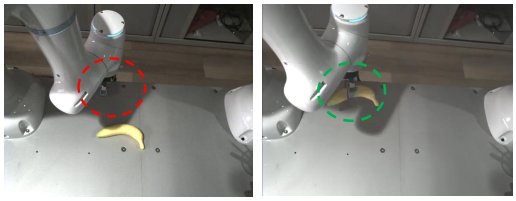

Figure 2: Left: Diff-CPL; Right: Diff-UAPA.

5.4 More Experiments

Ablation Studies. To assess the contributions of the diffusion policy and Beta prior components, we conducted ablation studies across multiple tasks. The evaluated variants include: 1) **BET-CPL**, which trains BET using the MLE objective, 2) **BET-UAPA**, which applies the MAP objective with a learned Beta prior, 3) **Diff-CPL**, which trains the diffusion policy with the MLE objective, and 4) **Diff-UAPA-Uniform**, which applies the MAP objective with a uniform Beta prior.

Table 3: Ablation results across Robomimic and Kitchen tasks. ✓ indicates the inclusion of the corresponding module, while △ indicates partial inclusion of the module.

Method	Modules		Robomimic				Kitchen			
	Diff	Beta	Lift	Can	Square	Transport	p1	p2	p3	p4
BET-CPL			49.2 ± 4.4	42.1 ± 1.1	57.6 ± 2.3	45.2 ± 4.8	97.0 ± 1.0	96.4 ± 0.5	88.4 ± 2.3	62.6 ± 2.0
BET-UAPA		✓	54.3 ± 0.5	48.0 ± 2.8	60.2 ± 2.5	52.3 ± 0.9	99.3 ± 0.9	99.3 ± 0.9	94.7 ± 1.0	68.0 ± 1.6
Diff-CPL	✓		48.6 ± 2.2	45.9 ± 2.8	55.2 ± 5.7	58.1 ± 6.2	**100.0 ± 0.0**	99.6 ± 0.2	94.2 ± 0.2	63.5 ± 0.8
Diff-UAPA-Uniform	✓	△	49.2 ± 0.7	48.5 ± 2.5	54.4 ± 2.7	58.2 ± 9.8	99.8 ± 0.3	98.3 ± 1.4	93.5 ± 2.9	64.1 ± 3.2
Diff-UAPA-C	✓	✓	**56.1 ± 0.9**	**61.3 ± 2.2**	**68.1 ± 0.6**	**64.0 ± 4.0**	**100.0 ± 0.0**	**99.7 ± 0.2**	**95.4 ± 0.6**	**70.9 ± 2.5**

The results are presented in Table 3. Comparing BET-based and Diff-based models highlights the superior performance of the diffusion policy. Additionally, evaluating methods with and without the Beta prior demonstrates the effectiveness of the proposed Beta model.

Various Noise Types. Following Bukharin et al. (2024), we experimented under two additional types of noises beyond the previously used irrational noise, including 1) *Stochastic Noise*, which generates preference labels by scaling the reward difference between pairs using a sampled temperature τ in (1,3) at each round; and 2) *Myopic Noise*, which derives preference labels from cumulative discounted rewards computed with an random discount factor γ in (0.5, 0.999) at each round.

We include a broader set of baselines for comparison. These include two-step PbRL approaches that first learn a reward function: 1) **BET-PbRL**, which integrates BET with TD3BC (Fujimoto & Gu, 2021), and 2) **Diff-PbRL**, which employs an efficient diffusion policy (Kang et al., 2023a) .Additionally, we consider robust preference alignment methods, including 4) **RIME** (Cheng et al., 2024) and 5) **R3M-DPO**(Bukharin et al., 2024). We use the Diff-UAPA-I variant for comparison.

Table 4: Performance comparison under different types of noisy preferences.

	HalfCheetah-Medium-Expert			HalfCheetah-Medium-Replay		
	Irrational Noise	Stochastic Noise	Myopic Noise	Irrational Noise	Stochastic Noise	Myopic Noise
BET-PbRL	2851 ± 45	3218 ± 102	2890 ± 81	1687 ± 91	1430 ± 90	1209 ± 45
Diff-PbRL	3158 ± 44	3298 ± 86	3002 ± 58	1730 ± 32	1499 ± 133	1389 ± 86
UA-PbRL	3291 ± 103	3410 ± 72	3198 ± 90	1803 ± 96	1674 ± 96	1503 ± 47
RIME	3200 ± 92	3397 ± 58	3090 ± 101	1891 ± 114	1592 ± 101	1459 ± 120
R3M-DPO	3258 ± 80	3505 ± 91	$\mathbf{3516 \pm 68}$	1914 ± 70	1643 ± 196	1655 ± 455
Diff-UAPA	$\mathbf{3297 \pm 101}$	$\mathbf{3674 \pm 169}$	3458 ± 118	$\mathbf{1949 \pm 53}$	$\mathbf{1715 \pm 57}$	$\mathbf{1707 \pm 85}$

The results shown in Table 4 demonstrate that the effectiveness of Diff-UAPA. This advantage stems from its ability to handle iterative preference alignment and progressively adapts the policy to these evolving signals. In contrast, baseline methods generally assume a static preference dataset, making them less effective in this dynamic setting. Moreover, robust preference learning methods that rely on filtering strategies may discard diverse yet informative preferences, thereby limiting the opportunity for comprehensive learning potentially useful preferences.

Noise Sensitivity. We also evaluate noise sensitivity in the Franka Kitchen environment by varying the reversal rate r from 50% (as used in previous experiments) to 25% and 75%, assessing robustness to label inconsistency. For clarity, we report only the most challenging p4 metric.

Table 5 presents the evaluation results. As the noise level increases (i.e., the reversal rate), all methods show a decline in performance, highlighting the significance of uncertainties in the dataset. However, compared to the other methods, Diff-UAPA consistently exhibits better performance with the highest success rate regardless of the scale of noise. This underscores the effectiveness of incorporating the Beta prior model to handle such uncertainties.

Table 5: Evaluation under different levels of reverse rates in the Kitchen environment.

	r=25%	r=50%	r=75%
BET-CPL	65.7 ± 1.6	62.6 ± 2.0	55.0 ± 2.5
BET-DPO	60.2 ± 4.8	57.4 ± 6.6	47.2 ± 7.0
Diff-CPL	66.0 ± 1.0	63.5 ± 0.8	57.1 ± 2.5
FKPD	71.3 ± 2.3	64.1 ± 3.2	62.3 ± 4.6
Diff-UAPA-C	75.3 ± 2.9	70.9 ± 2.5	$\mathbf{70.5 \pm 3.8}$
Diff-UAPA-I	$\mathbf{75.5 \pm 3.0}$	$\mathbf{71.7 \pm 4.6}$	69.1 ± 5.2

Reference Policy Sensitivity. To further verify the sensitivity to the success rate of the reference policy, we conducted a sensitivity test in the Robomimic environment on the Lift task by introducing reference policies with success rates of 25% and 75%.

From the results in Table 6, the success rate of the reference policy has a significant impact on the performance of the final policy. However, the performance of Diff-UAPA outperform than other algorithms in different success rates of reference policies, further demonstrating the importance of the beta prior component to the robustness of PbRL algorithms.

Table 6: Evaluation under different success rates of reference policies in Lift Task.

	r=25%	r=50%	r=75%
BET-CPL	40.7 ± 4.2	56.3 ± 3.4	79.0 ± 7.8
BET-DPO	37.3 ± 5.0	50.4 ± 3.3	76.3 ± 1.3
Diff-CPL	40.6 ± 5.7	53.7 ± 2.1	80.0 ± 2.0
FKPD	42.0 ± 2.0	56.2 ± 1.1	82.3 ± 1.3
Diff-UAPA-C	$\mathbf{44.7 \pm 4.1}$	$\mathbf{58.1 \pm 1.7}$	$\mathbf{84.0 \pm 3.5}$
Diff-UAPA-I	41.7 ± 2.3	56.1 ± 1.3	82.7 ± 3.1

6 Conclusion

In this paper, we present an uncertainty-aware preference alignment approach for diffusion policies using an iteratively updated preference dataset. Building on the maximum likelihood objective for directly aligning diffusion policies without learning a reward model, we introduce a Maximum A Posteriori (MAP) objective with an informative Beta prior, which is capable of capturing the uncertainty arising from potentially inconsistent human preferences. Empirical results across various domains demonstrate the effectiveness of our method. Future work will focus on extending this framework to the online setting, enabling agents to interact with the environment and dynamically adapt to evolving human preferences.

Acknowledgments

This work is supported in part by Shenzhen Science and Technology Major Program under grant KJZD20240903104008012, Shenzhen Fundamental Research Program (General Program) under grant JCYJ20230807114202005, Guangdong-Shenzhen Joint Research Fund under grant 2023A1515110617, CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund No. 2025A0505000053, Guangdong Basic and Applied Basic Research Foundation under grant 2024A1515012103, and Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

References

Ajay, A., Du, Y., Gupta, A., Tenenbaum, J., Jaakkola, T., and Agrawal, P. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations*, 2023.

An, G., Lee, J., Zuo, X., Kosaka, N., Kim, K.-M., and Song, H. O. Direct preference-based policy optimization without reward modeling. *Advances in Neural Information Processing Systems*, 36: 70247–70266, 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.

Bukharin, A., Hong, I., Jiang, H., Li, Z., Zhang, Q., Zhang, Z., and Zhao, T. Robust reinforcement learning from corrupted human feedback. *Advances in Neural Information Processing Systems*, 2024.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Chen, H., Lu, C., Ying, C., Su, H., and Zhu, J. Offline reinforcement learning via high-fidelity generative behavior modeling. In *The Eleventh International Conference on Learning Representations*, 2023a.

Chen, L., Bahl, S., and Pathak, D. Playfusion: Skill acquisition via diffusion from language-annotated play. In *Conference on Robot Learning*, pp. 2012–2029, 2023b.

Chen, Y., Li, H., and Zhao, D. Boosting continuous control with consistency policy. In *Autonomous Agents and Multiagent Systems*, pp. 335–344, 2024.

Cheng, J., Xiong, G., Dai, X., Miao, Q., Lv, Y., and Wang, F.-Y. Rime: Robust preference-based reinforcement learning with noisy preferences. In *International Conference on Machine Learning, ICML*, 2024.

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2023.

Choi, H., Jung, S., Ahn, H., and Moon, T. Listwise reward estimation for offline preference-based reinforcement learning. In *International Conference on Machine Learning*, 2024.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017.

Dong, Z., Yuan, Y., HAO, J., Ni, F., Mu, Y., ZHENG, Y., Hu, Y., Lv, T., Fan, C., and Hu, Z. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. In *International Conference on Learning Representations*, 2024.

Fang, B., Jia, S., Guo, D., Xu, M., Wen, S., and Sun, F. Survey of imitation learning for robotic manipulation. *Int. J. Intell. Robotics Appl.*, 3(4):362–369, 2019.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Gupta, A., Kumar, V., Lynch, C., Levine, S., and Hausman, K. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361, 2017.

Hejna, J. and Sadigh, D. Inverse preference learning: Preference-based rl without a reward function. *Advances in Neural Information Processing Systems*, 36, 2024.

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., and Sadigh, D. Contrastive preference learning: Learning from human feedback without rl. In *International Conference on Learning Representations*, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing systems*, 33:6840–6851, 2020.

Hwang, M., Lee, G., Kee, H., Kim, C. W., Lee, K., and Oh, S. Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 36:49088–49099, 2023.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. Reward learning from human preferences and demonstrations in atari. In *Advances in Neural Information Processing Systems, NeurIPS*, pp. 8022–8034, 2018.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022.

Joo, W., Lee, W., Park, S., and Moon, I. Dirichlet variational autoencoder. *Pattern Recognition*, 107: 107514, 2020.

Kang, B., Ma, X., Du, C., Pang, T., and Yan, S. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023a.

Kang, Y., Shi, D., Liu, J., He, L., and Wang, D. Beyond reward: Offline preference-guided policy optimization. In *International Conference on Machine Learning*, 2023b.

Kim, C., Park, J., Shin, J., Lee, H., Abbeel, P., and Lee, K. Preference transformer: Modeling human preferences using transformers for RL. In *International Conference on Learning Representations*, 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 preprint arXiv:2206.02231*, 2022.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*, 2021.

Liang, X., Shu, K., Lee, K., and Abbeel, P. Reward uncertainty for exploration in preference-based reinforcement learning. *arXiv preprint arXiv:2205.12401*, 2022.

Liu, B., Shao, S., Li, B., Bai, L., Xu, Z., Xiong, H., Kwok, J., Helal, S., and Xie, Z. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024.

Lu, C., Chen, H., Chen, J., Su, H., Li, C., and Zhu, J. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In *International Conference on Machine Learning*, pp. 22825–22855, 2023.

Mandal, D., Nika, A., Kamalaruban, P., Singla, A., and Radanović, G. Corruption robust offline reinforcement learning with human feedback. *arXiv preprint arXiv:2402.06734*, 2024.

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., and Martín-Martín, R. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Newman, M. E. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25, 2023.

Park, J., Seo, Y., Shin, J., Lee, H., Abbeel, P., and Lee, K. Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In *International Conference on Learning Representations*, 2022.

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 *Advances in Neural Information Processing Systems*, 2023.

Rafailov, R., Hejna, J., Park, R., and Finn, C. From r to q^*: Your language model is secretly a q-function. In *First Conference on Language Modeling*, 2024.

Shafiullah, N. M. M., Cui, Z. J., Altanzaya, A., and Pinto, L. Behavior transformers: Cloning k modes with one stone. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Shan, Z., Fan, C., Qiu, S., Shi, J., and Bai, C. Forward kl regularized preference optimization for aligning diffusion policies. *arXiv preprint arXiv:2409.05622*, 2024.

Shin, D., Dragan, A. D., and Brown, D. S. Benchmarks and algorithms for offline preference-based reward learning. *Transactions on Machine Learning Research*, 2023.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sridhar, A., Shah, D., Glossop, C., and Levine, S. Nomad: Goal masked diffusion policies for navigation and exploration. In *IEEE International Conference on Robotics and Automation*, pp. 63–70, 2024.

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

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Verma, M. and Metcalf, K. Hindsight priors for reward learning from human preferences. In *International Conference on Learning Representations*, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., Dai, B., and Miao, Q. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078, 2022.

Wang, Z., Hunt, J. J., and Zhou, M. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Warnell, G., Waytowich, N., Lawhern, V., and Stone, P. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Wirth, C., Akrour, R., Neumann, G., and Fürnkranz, J. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Xu, S., Yue, B., Zha, H., and Liu, G. A distributional approach to uncertainty-aware preference alignment using offline demonstrations. In *International Conference on Learning Representations*, 2025.

Xue, W., An, B., Yan, S., and Xu, Z. Reinforcement learning from diverse human preferences. In *International Joint Conference on Artificial Intelligence*, 2024.

Yang, L., Huang, Z., Lei, F., Zhong, Y., Yang, Y., Fang, C., Wen, S., Zhou, B., and Lin, Z. Policy representation via diffusion probability model for reinforcement learning. *arXiv preprint arXiv:2305.13122*, 2023.

Yuan, Y., Hao, J., Ma, Y., Dong, Z., Liang, H., Liu, J., Feng, Z., Zhao, K., and Zheng, Y. Uni-rlhf: Universal platform and benchmark suite for reinforcement learning with diverse human feedback. In *International Conference on Learning Representations, ICLR*, 2024.

Zhu, Z., Zhao, H., He, H., Zhong, Y., Zhang, S., Guo, H., Chen, T., and Zhang, W. Diffusion models for reinforcement learning: A survey. *arXiv preprint arXiv:2311.01223*, 2023.

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

Contents of Appendix

A More Details in Section 4.1

We detailed the deviation from Equation (10) to Equation (11) here.

$$
\begin{aligned}
&\mathcal{L}_{1,\text{MLE}}^{(\tau^w,\tau^l)}(\theta) \\
&= -\log\sigma\Big(\alpha\cdot\Big(\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,w}|s_t^w,a_t^{0,w})}\left[\gamma^t\log\frac{\pi_\theta(\overline{a_t^w}|s_t^w)}{\pi_{\text{ref}}(\overline{a_t^w}|s_t^w)}\right]-\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,l}|s_t^l,a_t^{0,l})}\left[\gamma^t\log\frac{\pi_\theta(\overline{a_t^l}|s_t^l)}{\pi_{\text{ref}}(\overline{a_t^l}|s_t^l)}\right]\Big)\Big) \\
&= -\log\sigma\Big(\alpha\cdot\Big(\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,\cdot}|s_t,a_t^{0,\cdot})}\left[\gamma^t\log\frac{\pi_\theta(\overline{a_t^w}|s_t^w)}{\pi_{\text{ref}}(\overline{a_t^w}|s_t^w)}-\gamma^t\log\frac{\pi_\theta(\overline{a_t^l}|s_t^l)}{\pi_{\text{ref}}(\overline{a_t^l}|s_t^l)}\right]\Big)\Big) \\
&= -\log\sigma\Big(\alpha\cdot\Big(\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,\cdot}|s_t,a_t^{0,\cdot})}\left[\sum_{i=1}^{I}\left(\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,l}|s_t^l))}{\pi_{\text{ref}}(a_t^{i-1|i,l}|s_t^l))}\right)\right]\Big)\Big) \\
&= -\log\sigma\Big(\alpha\cdot\Big(\mathbb{E}_{\pi_\theta(a_\cdot^{1:I,\cdot}|s_t,a_t^{0,\cdot})}\left[\sum_{t=0}^{T}\sum_{i=1}^{I}\left(\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,l}|s_t^l)}{\pi_{\text{ref}}(a_t^{i-1|i,l}|s_t^l)}\right)\right]\Big)\Big) \\
&= -\log\sigma\Big(\alpha I\cdot\Big(\mathbb{E}_{\substack{i\sim\mathcal{U}(0,I),\\ a_t^{i,w}\sim q(a_t^i|s_t^w,a_t^{0,w})\pi_\theta(a_t^{0,w}|s_t^w,a_t^{i,w}),\\ a_t^{i,l}\sim q(a_t^i|s_t^l,a_t^{0,l})\pi_\theta(a_t^{i-1,l}|s_t^l,a_t^{i,l})}}\left[\sum_{t=0}^{T}\left(\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,l}|s_t^l)}{\pi_{\text{ref}}(a_t^{i-1|i,l}|s_t^l)}\right)\right]\Big)\Big)
\end{aligned}
$$

Since $-\log\sigma(x)$ is a convex function:

$$
\left(-\log\sigma(x)\right)'' = \left(\sigma(x)-1\right)' = \left(\sigma(x)(1-\sigma(x))\right) \geqslant 0
$$

According to Jensen's inequality:

$$
\begin{aligned}
&\mathbb{E}_{\substack{i\sim\mathcal{U}(0,I),\\ a_t^{i,w}\sim q(a_t^{i,w}|a_t^{0,w},s_t^w),\\ a_t^{i,l}\sim q(a_t^{i,l}|a_t^{0,l},s_t^l)}}\Bigg[-\log\sigma\Bigg(\alpha I\cdot\mathbb{E}_{a_t^{i-1,\cdot}\sim\pi_\theta(\cdot)}\Bigg[\sum_{t=0}^{T}\Big(\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)} \\
&\qquad\qquad -\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,l}|s_t^l)}{\pi_{\text{ref}}(a_t^{i-1|i,l}|s_t^l)}\Big)\Bigg]\Bigg)\Bigg] \\
&= \mathbb{E}_{\substack{i\sim\mathcal{U}(0,I),\\ a_t^{i,w}\sim q(a_t^{i,w}|a_t^{0,w},s_t^w),\\ a_t^{i,l}\sim q(a_t^{i,l}|a_t^{0,l},s_t^l)}}\Bigg[-\log\sigma\Bigg(\alpha I\cdot\sum_{t=0}^{T}\Big(\gamma^t\mathbb{D}_{\text{KL}}\big[\pi_\theta(a_t^{i-1|i,w}|s_t^w)\;||\;\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)\big] \\
&\quad -\gamma^t\mathbb{D}_{\text{KL}}\big[\pi_\theta(a_t^{i-1|i,l}|s_t^l)\;||\;\pi_{\text{ref}}(a_t^{i-1|i,l}|s_t^l)\big]\Big)\Bigg)\Bigg]
\end{aligned}
\tag{18}
$$

According to Formula (1), it can be further simplified as:

$$
\begin{aligned}
-\mathbb{E}_{\substack{i\sim\mathcal{U}(0,I),\tau\in D_{pref},\\ a_t^\tau\sim q(a_t^\tau|a_t^0,s_t)}}\Bigg[\log\sigma\Bigg(-\alpha I\cdot\Bigg(&\sum_{t=0}^{T}\gamma^t(\|\epsilon^w-\epsilon_\theta(a_t^{\tau,w},s_t^w,\tau)\|_2^2-\|\epsilon^w-\epsilon_{ref}(a_t^{\tau,w},s_t^w,\tau)\|_2^2) \\
&-\sum_{t=0}^{T}\frac{\gamma^t}{|D_{pref}|}(\|\epsilon^\tau-\epsilon_\theta(a_t^{\tau,\tau},s_t^\tau,\tau)\|_2^2-\|\epsilon^\tau-\epsilon_{ref}(a_t^{\tau,\tau},s_t^\tau,\tau)\|_2^2)\Bigg)\Bigg)\Bigg]
\end{aligned}
$$

where 1) $i\sim\mathcal{U}(0,I)$ is the diffusion timestep, 2) $a_t^{i,w/l}\sim q(a_t^{i,w/l}|a_t^{0,w/l},s^{w/l})$ denotes the action $a_t^{0,w/l}$ corrupted with noise $\epsilon^{w/l}$ after i diffusion steps, and 3) $\epsilon_\theta^{w/l}$ is the noise predictor.

B More Details in Section 4.2

We detailed the deviation of Equation (15) here.

$$
\begin{aligned}
&\mathcal{L}_{1,\text{prior}}^{(\tau^w,\tau^l)}(\theta)\\
&= -\log\sigma\Bigg(\text{Beta}\Bigg(\alpha\cdot\Bigg(\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,w}|s_t^w,a_t^{0,w})}\left[\gamma^t\log\frac{\pi_\theta(\overline{a_t^w}|s_t^w)}{\pi_{\text{ref}}(\overline{a_t^w}|s_t^w)}\right]-\sum_{\tau\in\mathcal{D}_{\text{pref}}}\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,\tau}|s_t^\tau,a_t^{0,\tau})}\left[\gamma^t\log\frac{\pi_\theta(\overline{a_t^\tau}|s_t^\tau)}{\pi_{\text{ref}}(\overline{a_t^\tau}|s_t^\tau)}\right]\Bigg);\alpha+1,\beta+1\Bigg)\Bigg)\\
&= -\log\sigma\Bigg(\text{Beta}\Bigg(\alpha\cdot\Bigg(\mathbb{E}_{\pi_\theta(a_t^{1:I,\cdot}|s_t^\cdot,a_t^{0,\cdot})}\left[\sum_{t=0}^{T}\gamma^t\log\frac{\pi_\theta(\overline{a_t^w}|s_t^w)}{\pi_{\text{ref}}(\overline{a_t^w}|s_t^w)}-\sum_{\tau\in\mathcal{D}_{\text{pref}}}\sum_{t=0}^{T}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}\log\frac{\pi_\theta(\overline{a_t^\tau}|s_t^\tau)}{\pi_{\text{ref}}(\overline{a_t^\tau}|s_t^\tau)}\right]\Bigg);\alpha+1,\beta+1\Bigg)\Bigg)\\
&= -\log\sigma\Bigg(\alpha\cdot\Bigg(\sum_{\tau\in\mathcal{D}_{\text{pref}}}\sum_{t=0}^{T}\mathbb{E}_{\pi_\theta(a_t^{1:I,\cdot}|s_t^\cdot,a_t^{0,\cdot})}\left[\sum_{i=1}^{I}\left(\sum_{t=0}^{T}\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\sum_{\tau\in\mathcal{D}_{\text{pref}}}\sum_{t=0}^{T}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}\log\frac{\pi_\theta(a_t^{i-1|i,\tau}|s_t^\tau)}{\pi_{\text{ref}}(a_t^{i-1|i,\tau}|s_t^\tau)}\right)\right]\Bigg)\Bigg)\\
&= -\log\sigma\Bigg(\alpha\cdot\Bigg(\mathbb{E}_{\pi_\theta(a_\cdot^{1:I,\cdot}|s_t^\cdot,a_t^{0,\cdot})}\left[\sum_{t=0}^{T}\sum_{i=1}^{I}\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\sum_{\tau\in\mathcal{D}_{\text{pref}}}\sum_{t=0}^{T}\sum_{i=1}^{I}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}\log\frac{\pi_\theta(a_t^{i-1|i,\tau}|s_t^\tau)}{\pi_{\text{ref}}(a_t^{i-1|i,\tau}|s_t^\tau)}\right]\Bigg)\Bigg)\\
&= -\log\sigma\Bigg(\alpha I\cdot\Bigg(\mathbb{E}_{\substack{i\sim\mathcal{U}(0,I),\\ a_t^{i,\cdot}\sim q(a_t^i|s_t^\cdot,a_t^{0,\cdot})\\ \pi_\theta(a_t^{0,\cdot}|s_t^\cdot,a_t^{i,\cdot})}}\left[\sum_{t=0}^{T}\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\sum_{\tau\in\mathcal{D}_{\text{pref}}}\sum_{t=0}^{T}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}\log\frac{\pi_\theta(a_t^{i-1|i,\tau}|s_t^\tau)}{\pi_{\text{ref}}(a_t^{i-1|i,\tau}|s_t^\tau)}\right]\Bigg)\Bigg).
\end{aligned}
$$

Since $-\log\sigma(\text{Beta}(x;\alpha,\beta))$ is a convex function when $\alpha+\beta\geqslant 2$. Define $g(t)=-\log\big(\sigma(t)\big)$. Since

$$-\log\big(\sigma(t)\big)=\log\big(1+e^{-t}\big),$$

it suffices to show that $\log(1+e^{-t})$ is convex in t. Differentiating,

$$\frac{d}{dt}\log\big(1+e^{-t}\big)=\frac{-e^{-t}}{1+e^{-t}}=-\frac{1}{e^t+1},$$

and hence

$$\frac{d^2}{dt^2}\log\big(1+e^{-t}\big)=\frac{e^t}{(e^t+1)^2}\;>\;0\quad(\forall t\in\mathbb{R}).$$

This shows $\log(1+e^{-t})$ is strictly convex in t. Therefore, for the function

$$f(x)=-\log\Big[\sigma\big(\text{Beta}(x;\,\alpha+1,\beta+1)\big)\Big],$$

the inner part $\text{Beta}(x;\,\alpha+1,\beta+1)$ serves as the real argument t, and the composition preserves convexity, implying $f(x)$ is convex.
According to Jensen's inequality

$$
\begin{aligned}
&\mathbb{E}_{\substack{\tau\in\mathcal{D}_{\text{pref}},\\ a_t^{i,\cdot}\sim q(a_t^{i,\cdot}|a_t^{0,\cdot},s_t^\cdot)}}\Bigg[-\log\sigma\Big(\text{Beta}\Big(\alpha I\cdot\Big(\mathbb{E}_{a_t^{i-1,\cdot}\sim\pi_\theta(a_t^{i-1,\cdot}|s_t^\cdot,a_t^{0,\cdot})}\\
&\Big[\Big(\sum_{t=0}^{T}\gamma^t\log\frac{\pi_\theta(a_t^{i-1|i,w}|s_t^w)}{\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)}-\sum_{\tau\in\mathcal{D}_{\text{pref}},t=0}^{T}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}\log\frac{\pi_\theta(a_t^{i-1|i,\tau}|s_t^\tau)}{\pi_{\text{ref}}(a_t^{i-1|i,\tau}|s_t^\tau)}\Big)\Big]\Big);\alpha+1,\beta+1\Big)\Big)\Bigg]\\
&=\mathbb{E}_{\substack{\tau\in\mathcal{D}_{\text{pref}},\\ a_t^{i,\cdot}\sim q(a_t^{i,\cdot}|a_t^{0,\cdot},s_t^\cdot)}}\Bigg[-\log\sigma\Big(\alpha I\cdot\sum_{t=0}^{T}\Big(\gamma^t\mathbb{D}_{\text{KL}}\Big[\pi_\theta(a_t^{i-1|i,w}|s_t^w)\;||\;\pi_{\text{ref}}(a_t^{i-1|i,w}|s_t^w)\Big]\\
&-\sum_{\tau\in\mathcal{D}_{\text{pref}},t=0}^{T}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}\mathbb{D}_{\text{KL}}\Big[\pi_\theta(a_t^{i-1|i,\tau}|s_t^\tau)\;||\;\pi_{\text{ref}}(a_t^{i-1|i,\tau}|s_t^\tau)\Big]\Big)\Big)\Bigg]
\end{aligned}
$$

According to Formula (1), it can be further simplified as:

$$
\begin{aligned}
-\mathbb{E}_{\substack{\tau\in\mathcal{D}_{\text{pref}},\\ a_t^{i,\cdot}\sim q(a_t^{i,\cdot}|a_t^{0,\cdot},s_t^\cdot)}}\Bigg[\log\sigma\Big(-\alpha I\cdot\Big(\sum_{t=0}^{T}\gamma^t(\|\epsilon^w-\epsilon_\theta(a_t^{i,w},s_t^w,i)\|_2^2-\|\epsilon^w-\epsilon_{\text{ref}}(a_t^{n,w},s_t^w,i)\|_2^2)\\
-\sum_{\tau\in\mathcal{D}_{\text{pref}},t=0}^{T}\frac{\gamma^t}{|\mathcal{D}_{\text{pref}}|}(\|\epsilon^\tau-\epsilon_\theta(a_t^{i,\tau},s_t^\tau,i)\|_2^2-\|\epsilon^\tau-\epsilon_{\text{ref}}(a_t^{i,\tau},s_t^\tau,i)\|^2)\Big)\Big)\Bigg]
\end{aligned}
$$

where 1) $i \sim \mathcal{U}(0, I)$ is the diffusion timestep, 2) $a_t^{i,\cdot} \sim q(a_t^{i,\cdot}|a_t^{0,\cdot}, s^{\cdot})$ denotes the action $a_t^{0,\cdot}$ corrupted with noise $\epsilon^{\cdot}$ after i diffusion steps, and 3) $\epsilon_\theta^{\cdot}$ is the noise predictor.

C Proof of Proposition 4.1

Proposition 4.1 can be divided into two parts: 1) the uncertainty-aware property of the Beta prior, and 2) the prior on the strength of a trajectory.

Part 1. We show the uncertainty-aware capability of the Beta prior $\text{Beta}(\phi(\tau); \alpha, \beta)$ during the iterative preference alignment process outlined in Definition 3.1 as follows.

The probability density function (PDF) of the Beta distribution $\text{Beta}(\phi(\tau); \alpha, \beta)$ is given by:

$$f(\phi(\tau); \alpha, \beta) = \frac{\phi(\tau)^{\alpha-1}(1-\phi(\tau))^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leqslant \phi(\tau) \leqslant 1, \tag{19}$$

where $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt$ is the Beta function, serving as a normalizing constant.

The variance of a Beta distribution $\text{Beta}(\phi(\tau); \alpha, \beta)$ is given by the following formula:

$$\text{Var}(\text{Beta}(\alpha, \beta)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{20}$$

In the process described in Definition 3.1, the uncertainty arises from the varying preferences of different human raters for a given trajectory pair (τ^i, τ^j). Without loss of generality, assuming an initial belief of $\text{Beta}(1, 1)$ for each trajectory, and with 10 raters evaluating a candidate pair (τ^i, τ^j), the Beta prior is updated according to the preferences expressed by the raters. For instance, in the first case, where 9 raters prefer τ^i and 1 rater prefers τ^j, the Beta prior for τ^i would be updated to $\text{Beta}(10, 2)$. In the second case, where 5 raters prefer τ^i and 5 prefer τ^j, the Beta prior for τ^i would become $\text{Beta}(6, 6)$. Intuitively, we would be more confident with less uncertainty in the first case, as the majority of raters share the same preference.

The Beta distribution effectively captures this uncertainty. As shown in Equation (20), the variance of $\text{Beta}(10, 2)$ is smaller than that $\text{Beta}(6, 6)$, indicating that $\text{Beta}(10, 2)$ is 'sharper' and reflects less uncertainty, which aligns with our intuition.

Part 2. We prove that the prior on the strength of a trajectory is proportional to $\text{Beta}((\phi(\tau); \alpha + 1, \beta + 1))$, i.e., $p_0(A^{\pi_\theta}(\tau)) \propto \text{Beta}(\phi(\tau); \alpha + 1, \beta + 1)$, as follows.

Recall that the probability of a trajectory τ with strength $A^{\pi_\theta}(\tau)$ winning against the average candidate is given by $\phi(\tau) = \sigma(A^{\pi_\theta}(\tau) - \bar{A}^{\pi_\theta}) \in (0, 1)$. Let $A^{\pi_\theta}(\tau) - \bar{A}^{\pi_\theta}$ be denoted as $\tilde{A}^{\pi_\theta}(\tau)$. According to Equation (19), we have that the Beta distribution over $\phi(\tau) = \sigma(\tilde{A}^{\pi_\theta}(\tau))$ is:

$$\text{Beta}(\sigma(\tilde{A}^{\pi_\theta}(\tau)); \alpha, \beta) \propto \sigma(\tilde{A}^{\pi_\theta}(\tau))^{\alpha-1}(1 - \sigma(\tilde{A}^{\pi_\theta}(\tau)))^{\beta-1}. \tag{21}$$

The derivative of the sigmoid function is:

$$\sigma'(\tilde{A}^{\pi_\theta}(\tau)) = \sigma(\tilde{A}^{\pi_\theta}(\tau))(1 - \sigma(\tilde{A}^{\pi_\theta}(\tau))). \tag{22}$$

By incorporating Equation (21) and Equation (22) into Equation (14), we have that:

$$\begin{aligned} p_0(A^{\pi_\theta}(\tau)) &\propto \sigma(\tilde{A}^{\pi_\theta}(\tau))^{\alpha}(1 - \sigma(\tilde{A}^{\pi_\theta}(\tau)))^{\beta} \\ &\propto \text{Beta}(\sigma(\tilde{A}^{\pi_\theta}(\tau)); \alpha + 1, \beta + 1) \\ &= \text{Beta}(\phi(\tau); \alpha + 1, \beta + 1). \end{aligned} \tag{23}$$

D More Experimental Details

D.1 Experimental Settings

In this paper, we utilized a total of 4 NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory. The random seeds used for the experiments were 42, 43, and 44. Each experiment is repeated

using these random seeds, and the mean $\pm$ standard deviation (std) of the results is reported. The learning rate is reset at the beginning of each round to enhance stability. We trained the agents offline and selected the final epoch for evaluation across 56 parallel environments, each with 10 episodes. Additionally, we employed a transformer-based architecture for the Beta model as in the preference transformer (Kim et al., 2023).

D.2 Environmental Details

Manipulation Tasks. Robomimic (Mandlekar et al., 2021) is a large-scale robotic manipulation benchmark designed to explore imitation learning and offline reinforcement learning (RL). It consists of five tasks, each with a proficient human (PH) teleoperated demonstration dataset, and four tasks also feature mixed proficient/non-proficient human (MH) demonstration datasets, resulting in a total of nine variants. In this paper, we focus on three tasks: Lift, Can, and Square. Specifically:

- Lift: The robot arm must lift a small cube. This is the simplest task.
- Can: The robot must move a Coke can from a large bin to a smaller target bin. This task is slightly more challenging than Lift, as picking up the can is more difficult than picking up the cube, and the can must be placed accurately in the target bin.
- Square: The robot is required to pick up a square nut and place it onto a rod. This task is significantly more difficult than Lift and Can, as it demands high precision to pick up the nut and insert it into the rod.
- Transport: The robot needs to open the box and transport the hammer in the box to another robot.

The Franka Kitchen is also a widely used environment for evaluating the performance of methods in learning complex, long-horizon tasks. Introduced in Relay Policy Learning (Gupta et al., 2019), the environment features seven objects for interaction and includes a human demonstration dataset consisting of 566 demonstrations, each completing four tasks in random order. The objective is to execute as many of the demonstrated tasks as possible, regardless of their order, highlighting both short-horizon and long-horizon multimodal capabilities.

Locomotion Tasks. We evaluate our locomotion tasks using the D4RL benchmark (Fu et al., 2020), which is widely used in reinforcement learning (RL) for continuous control tasks. In this paper, we focus on the Hopper and HalfCheetah environments. In these environments, the goal is to maximize the cumulative reward within a single episode by navigating a sequence of actions that optimize the agent's movement and efficiency. More specifically:

- Hopper: In this task, the agent controls a 2D hopping robot, with the objective of balancing and moving the robot forward using as few steps as possible.
- HalfCheetah: In this task, the agent controls a 2D robotic cheetah, aiming to run as fast as possible while maximizing speed and maintaining stability.

We use the medium-expert and medium-replay datasets for both environments. The the medium-expert dataset combines expert demonstrations with suboptimal trajectories, while the medium-replay dataset contains the replay buffer from a partially trained SAC policy (Haarnoja et al., 2017).

D.3 Computational Overhead

The additional computational overhead can primarily be attributed to the following components:

Diffusion policy. While diffusion policies incur higher computational costs than simpler architectures like MLPs, this overhead is partially offset by the action sequence prediction strategy in (Chi et al., 2023). More importantly, diffusion models are widely adopted in RL for their strong generative capabilities and superior performance. In practice, training time for diffusion is roughly twice that of the transformer in our experiments.

Beta model. In this work, we use efficient techniques like the reparameterization trick to improve scalability. In practice, the computational cost of training the Beta model is similar to training a reward model in traditional PbRL. Since our method avoids training a reward model, the added cost is less effective compared to conventional PbRL. Additionally, the extra computational cost only slightly increases training time—by a few minutes—while the subsequent RL phase is much more demanding, often taking several hours.

D.4 Manipulation Preference Dataset

For the robot manipulation tasks, we train two policies using behavior cloning: the BET policy and the diffusion policy. Training proceeds until a 40% success rate is reached. To build the simulation environment, we deploy 56 parallel environments, each initialized with a different seed to ensure varied initial positions for the agent. We then collect 560 trajectories per policy. From these, we randomly select 500 trajectory pairs and label them based on the sum of their rewards. During training, each trajectory is sliced using the observed steps as the stride, and these segments are compared. In the iterative update process, for each update round, we randomly select 20% of the trajectory pairs and apply a 50% reversal rate by swapping the winner and loser. To improve stability and convergence, the learning rate is reset at the start of each round.

D.5 Hyperparameters

Our experiments are primarily based on the codebase from (Chi et al., 2023). Therefore, we retain the same hyperparameters for training the diffusion policy as specified in (Chi et al., 2023) for each experiment. The specific hyperparameters for Diff-UAPA are listed in Table 7.

Table 7: List of the specific hyperparameters for the proposed Diff-UAPA. To ensure fair comparisons, we maintain consistency in other parameters of the same neural networks across different models.

Parameters	Robomimic	Kitchen	D4RL
General			
Training Epochs	600	600	600
Episode Length	400	280	1000
Beta Model			
Network	256	256	256
Learning Rate	2e-5	2e-5	3e-5
Number of Attention Heads	4	4	4
Number of Layers	2	2	1
Batch Size	32	32	64
Initial Belief	$\alpha = \beta = 1$	$\alpha = \beta = 1$	$\alpha = \beta = 1$

D.6 Visualization Results

Figure 3 presents visualization results from the manipulation tasks. It is evident that the baseline method, Diff-CPL, which is trained using the MLE objective, struggles to handle certain critical scenarios, particularly those involving noisy preferences.

Figure 4 presents visualization results from the real-world pick-and-place experiment. As shown, Diff-UAPA successfully completes the task by picking up and placing the banana, whereas Diff-CPL struggles to pick up the banana due to the impact of learning from noisy preference labels.

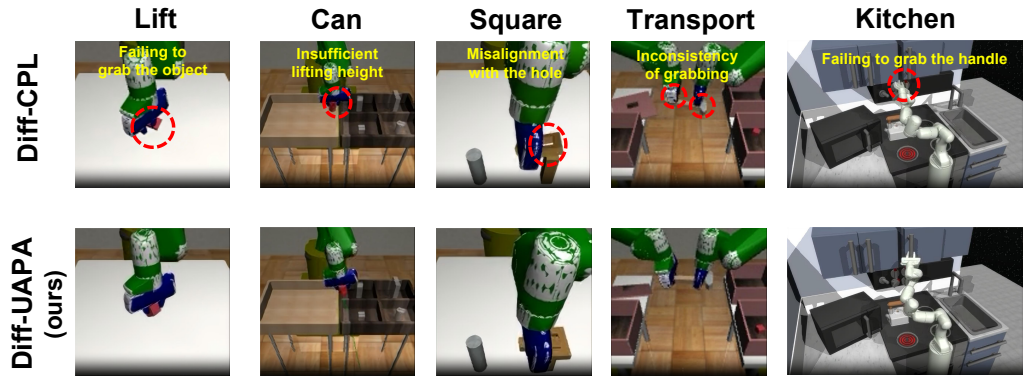

Figure 3: Visualization results in five manipulation tasks.

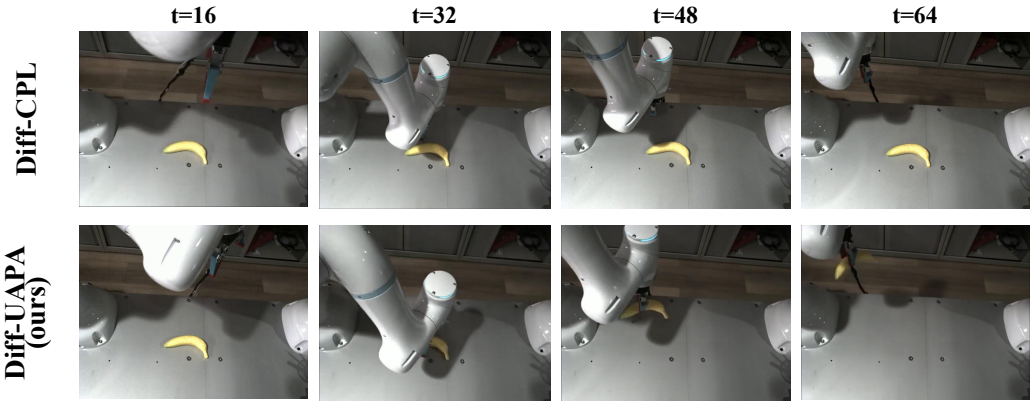

Figure 4: Visualization results in the pick-and-place task.

E Limitation

Offline Trajectory Dataset. This paper primarily focuses on learning from an offline trajectory dataset with potentially inconsistent human preferences that are iteratively updated, where the agent cannot directly interact with the environment. This partial offline setup may limit the agent's ability to explore and discover improved strategies through interactive online learning. However, our method can also generalize to an online setting, where both trajectories and human preferences are dynamically updated over time.

Computational Overhead. The integration of training a Beta prior model through variational inference adds computational complexity compared to simpler MLE-based methods. However, by utilizing efficient techniques like the reparameterization trick to enhance scalability, the computational overhead of training the Beta model is minimal in practice, adding only a small additional time cost relative to the diffusion training process.

F Social Impact

The proposed Diff-UAPA framework presents meaningful implications for aligning AI agents with diverse and potentially conflicting human preferences. By explicitly modeling uncertainty through a Bayesian prior, this approach promotes fairness and inclusivity in decision-making systems by preventing the marginalization of minority or inconsistent viewpoints. In high-stakes domains such as assistive robotics, healthcare, or automated systems interacting with vulnerable populations, Diff-UAPA's robustness to noisy and heterogeneous feedback helps ensure safer and more equitable outcomes. However, care must be taken in interpreting preference data, as biases in human feedback can propagate through the model. As with any alignment technique, ethical considerations related to whose preferences are prioritized and how disagreements are resolved remain critical areas.

G Author Contributions

Sheng Xu and Guiliang Liu conceived the initial ideas, organized, and led the research project. Runqing Miao developed the code, conducted all the experiments, and prepared the initial manuscript draft. Sheng Xu contributed to the experimental design and refined and finalized the writing. Runyi Zhao assisted with the real-world experiments. Wai Kin Victor Chan provided valuable feedback, and Guiliang Liu offered overall supervision and constructive guidance throughout the study. Runqing Miao and Sheng Xu are co-first authors with equal contributions, listed in alphabetical order.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: They can support the main claims of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section E.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We rigorously prove our theoretical results, and the details can be found in Section A, Section B, and Section C.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We report our experimental settings and hyperparameters in Section D, and provide our code in the Supplemental Materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide our code in the Supplemental Materials with a clear README file to reproduce the results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We report our experimental settings and hyperparameters in Section D, and provide our code in the Supplemental Materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We perform the experiments under three random seeds and report the mean and std results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute resources in Section D.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Section F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We strictly follow the corresponding licenses for existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We only use the simulated and public data, without involving crowd-sourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.