

TOWARDS BETTER OPTIMIZATION FOR LISTWISE PREFERENCE IN DIFFUSION MODELS

Jiamu Bai^{1,*}, Xin Yu^{1,*}, Meilong Xu³, Weitao Lu², Xin Pan²,
Kiwan Maeng¹, Daniel Kifer¹, Jian Wang², Yu Wang²

¹Penn State University ²TikTok Inc. ³Stony Brook University



Figure 1: Sample images generated from SDXL trained with Diffusion-LPO. Diffusion-LPO generalizes Diffusion-DPO by optimizing the preference under a list of ranked images. After finetuning with Diffusion-LPO, SDXL produces images with higher visual aesthetics and prompt alignments.

ABSTRACT

Reinforcement learning from human feedback (RLHF) has proven effectiveness for aligning text-to-image (T2I) diffusion models with human preferences. Although Direct Preference Optimization (DPO) is widely adopted for its computational efficiency and avoidance of explicit reward modeling, its applications to diffusion models have primarily relied on pairwise preferences. The precise optimization of listwise preferences remains largely unaddressed. In practice, human feedback on image preferences often contains implicit ranked information, which

* Co-first author. Work done during Jiamu Bai’s internship at TikTok.

1:{jvb6867,xmy5152,kvm6242,duk17}@psu.edu

2:{weitao.lu,xin.pan,jian.wang1,yuwang.w}@bytedance.com

3:meixu@cs.stonybrook.edu

conveys more precise human preferences than pairwise comparisons. In this work, we propose **Diffusion-LPO**, a simple and effective framework for **Listwise Preference Optimization** in diffusion models with listwise data. Given a caption, we aggregate user feedback into a ranked list of images and derive a listwise extension of the DPO objective under the Plackett–Luce model. Diffusion-LPO enforces consistency across the entire ranking by encouraging each sample to be preferred over all of its lower-ranked alternatives. We empirically demonstrate the effectiveness of Diffusion-LPO across various tasks, including text-to-image generation, image editing, and personalized preference alignment. Diffusion-LPO consistently outperforms pairwise DPO baselines on visual quality and preference alignment.

1 INTRODUCTION

Diffusion models, including prominent architectures like Stable Diffusion (Podell et al., 2023; Rombach et al., 2022), Imagen (Saharia et al., 2022), and DALL-E 2 (Ramesh et al., 2022), have become a dominant paradigm in text-to-image synthesis due to their ability to generate high-fidelity and semantically aligned images (Ho et al., 2020). While these models are pretrained on massive web-scale datasets to establish a powerful foundation, real-world deployment necessitates further finetuning to align their outputs more closely with human preferences. Drawing inspiration from alignment techniques in large language models (LLMs), recent works have adapted methods like Direct Preference Optimization (DPO) for diffusion models, enabling preference learning without an explicit reward model (Yang et al., 2024; Wallace et al., 2024; Zhu et al., 2025; Li et al., 2024; Gu et al., 2024). These approaches have demonstrated superior performance over standard supervised fine-tuning by directly incorporating preference signals into the optimization process.

Despite these advances, prior works mostly rely on pairwise human preference data. Pairwise comparisons are relatively easy to collect and annotate, making them the dominant form in existing human preference datasets (Bai et al., 2022; Kirstain et al., 2023). However, human preferences are inherently expressed as ranked lists rather than isolated pairs, since

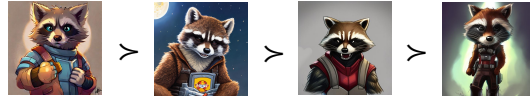


Figure 2: An example of a ranked list of images under human preference. The caption is “Rocket Raccoon, furry art, fanart, digital painting.”

preferences are not simply absolute approvals or rejections but relative orderings among multiple options, as shown in the example of Figure 2. In this sense, pairwise annotations implicitly contain richer ranking information: if we consider a user indicating their preferences as response $\mathbf{x}^{(a)} \succ \mathbf{x}^{(b)}$ and $\mathbf{x}^{(b)} \succ \mathbf{x}^{(c)}$, this naturally implies the transitive ordering $\mathbf{x}^{(a)} \succ \mathbf{x}^{(b)} \succ \mathbf{x}^{(c)}$, which more faithfully represents the user’s overall preference. Notably, we observe that in the Pick-a-Pic dataset (Kirstain et al., 2023), where each user provides pairwise preferences for images, 56% of such annotations can be benefited from grouping into consistent rankings with size greater than 2 to form more informative human preferences. This motivates the need for listwise optimization, which directly models the ranking nature of human preferences. A straightforward approach to formulating an objective for listwise preference is to decompose rankings of size m into $m(m-1)/2$ pairs to apply the pairwise DPO objective. Several works (Chen et al., 2025; Karthik et al., 2024) have explored this direction, but their methods rely on auxiliary evaluators to assign reward scores to each image, introducing extra computational overhead and resource burden.

In this work, we introduce Diffusion-LPO (Listwise Preference Optimization), a new framework for aligning diffusion models with user-level preferences. Instead of reducing user feedback to pairwise comparisons, Diffusion-LPO directly models listwise preferences with the Plackett-Luce ranking model (Plackett, 1975). Unlike pairwise DPO, which only compares a winning sample against a single loser, Diffusion-LPO enforces consistency across the entire ranking by encouraging each sample to be preferred over all of its lower-ranked alternatives, thereby preserving the full relative order within the list. When the list size equals 2, Diffusion-LPO degenerates to Diffusion-DPO. As a general approach for listwise preference optimization, the formulation of Diffusion-LPO can potentially enhance many existing methods built upon the Diffusion-DPO family.

To evaluate the effectiveness of our method, we train SD1.5 (Rombach et al., 2022) and SDXL (Podell et al., 2023) on the Pick-a-Pic dataset (Kirstain et al., 2023), a dataset of real-world user preferences that contains human preference pairs of images. We uncover the inherent listwise preference within the dataset, with detailed information of dataset reformulation illustrated in Section 4.1. Empirical results show that Diffusion-LPO improves PickScore win rates by 12% and by 4% over Diffusion-DPO on SD1.5 and SDXL. Specifically, our contribution can be listed as follows:

- We first provide a novel perspective that ranked human preferences are implicitly embedded within pairwise annotations. It motivates us to intentionally enforce the diffusion model to learn such structural rank information.
- We propose Diffusion-LPO, a direct preference optimization framework from listwise human feedback. As a generalization of pairwise DPO, Diffusion-LPO can be integrated into any DPO-based method to further boost the performance.
- We empirically demonstrate the effectiveness of Diffusion-LPO across various T2I tasks. Diffusion-LPO exhibits higher alignment with human preferences over baselines.

2 RELATED WORK

RLHF Alignment Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant paradigm for aligning large language models (LLMs) with human preferences (Christiano et al., 2017; Stiennon et al., 2020; Mnih et al., 2016; Ziegler et al., 2019; Bai et al., 2022; Bakker et al., 2022; Jiang et al., 2025). In its standard formulation, a reward model is trained to approximate user preferences, and the policy is subsequently optimized via reinforcement learning with this reward signal. While effective, this two-stage pipeline is computationally expensive and can be susceptible to reward hacking (Singhal et al., 2023; Chen et al., 2024). To address these limitations, recent work has explored direct preference optimization (DPO) (Rafailov et al., 2023), which bypasses explicit reward modeling and instead optimizes the policy to prefer human-preferred outputs directly. Most existing approaches rely on pairwise preference data and the Bradley-Terry (BT) model (Bradley & Terry, 1952) to parameterize the likelihood of a winning sample over a losing one (Ethayarajh et al., 2024; Meng et al., 2024). However, pairwise comparisons may provide limited information about human preferences, as they reduce complex judgments into binary wins and losses. Several recent works have extended DPO to the listwise ranking setting (Liu et al., 2025; Song et al., 2024; Hong et al., 2024a), enabling richer supervision signals from ranked lists of responses than pairs.

Preference Alignment with Diffusion Models Recent work has shown that reinforcement learning is an effective tool for aligning diffusion models with human aesthetic and semantic preferences (Fan et al., 2023; Black et al., 2023; Xue et al., 2025; Clark et al., 2023; Zhao et al., 2025; Uehara et al., 2024; Prabhudesai et al., 2023; Lee et al., 2023; Eyring et al., 2024; Sun et al., 2025a; Lee et al., 2025; Zhang et al., 2025). Among existing reinforcement learning optimization approaches, Direct Preference Optimization (DPO) stands out as an attractive option by defining an implicit reward through preference comparisons, eliminating the need for explicit reward models (Wallace et al., 2024; Yang et al., 2024; Li et al., 2024; Gu et al., 2024; Zhu et al., 2025; Hong et al., 2024b; Lee et al., 2025; Wang et al., 2025; Liang et al., 2024; Sun et al., 2025b; Lu et al., 2025; Wu et al., 2025; Cai et al., 2025; Hu et al., 2025; Zhang et al., 2024). Diffusion-DPO (Wallace et al., 2024) adapts this framework to the diffusion setting. Follow-up work includes DSPO (Zhu et al., 2025) which derives the DPO objective through score matching, Diffusion-KTO (Li et al., 2024) which formulates the objective by maximizing human utility, MaPO (Hong et al., 2024b) which maximizes likelihood margin between preferred and dispreferred image sets. Despite these advances, the above DPO-based methods for diffusion rely exclusively on *pairwise* preference data, while human feedback can come as a list of rankings.

3 PRELIMINARIES

3.1 DIFFUSION MODELS

Diffusion models are generative models that learn data distributions by reversing a gradual noising process (Ho et al., 2020; Song et al., 2021). The forward (diffusion) process starts from clean data $x_0 \sim p_{\text{data}}$, where p_{data} is the data distribution, and progressively adds Gaussian noise across T time steps with noise scheduling $\beta_1, \dots, \beta_T$:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t)I), \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s, \quad \alpha_t = 1 - \beta_t.$$

The reverse process (denoising) is parameterized by a neural network ϵ_θ , which predicts noise ϵ_t given a noisy image $\mathbf{x}_t$ at timestep t . The standard training objective is a weighted denoising score matching loss (Ho et al., 2020):

$$\mathcal{L}_{\text{DM}}(\theta) = \mathbb{E}_{\mathbf{x}_t^{(1:m)} \sim p_\theta(\mathbf{x}_t^{(1:m)} | \mathbf{x}_0^{(1:m)})} [\omega(\lambda_t) \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2],$$

where $\lambda_t = \log(\alpha_t^2/\sigma_t^2)$ is the signal-to-noise ratio, p_θ is the reverse process, and $\omega(\lambda_t)$ a pre-defined weighting function.

3.2 DIRECT PREFERENCE OPTIMIZATION (DPO)

Large generative models can be aligned with human feedback via reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020). The goal of RLHF is to optimize the policy π_θ by maximizing the reward $r(\mathbf{c}, \mathbf{x}_0)$ given the prompt $\mathbf{c}$ sampled from prompt set $\mathcal{D}_c$ and sample $\mathbf{x}_0$, with a regularization KL-divergence from a reference policy $\pi_{\text{ref}}(\mathbf{x}_0 | \mathbf{c})$:

$$\max_{\pi_\theta} \mathbb{E}_{\mathbf{c} \sim \mathcal{D}_c, \mathbf{x}_0 \sim \pi_\theta(\mathbf{x}_0 | \mathbf{c})} [r(\mathbf{c}, \mathbf{x}_0)] - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(\mathbf{x}_0 | \mathbf{c}) || \pi_{\text{ref}}(\mathbf{x}_0 | \mathbf{c})]. \quad (1)$$

Classical RLHF requires training a separate reward model and then optimizing the generative policy with reinforcement learning (e.g., PPO (Schulman et al., 2017)). Direct Preference Optimization (DPO) (Rafailov et al., 2023) eliminates the need for explicit reward modeling by modeling human preferences with the Bradley-Terry (BT) model (Bradley & Terry, 1952):

$$p_{BT}(\mathbf{x}_0^+ \succ \mathbf{x}_0^- | \mathbf{c}) = \sigma(r(\mathbf{c}, \mathbf{x}_0^+) - r(\mathbf{c}, \mathbf{x}_0^-)),$$

where $\mathbf{c}$ is the conditioning prompts and $(\mathbf{x}_0^+, \mathbf{x}_0^-)$ is the pair of human-labeled winning and losing samples. Here, $r(\mathbf{c}, \mathbf{x}_0)$ is the latent reward, and σ is the sigmoid function. Use the log-likelihood ratios to get the implicit rewards, the DPO loss becomes

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \left(\log \pi_\theta(\mathbf{x}_0^+ | \mathbf{c}) - \log \pi_\theta(\mathbf{x}_0^- | \mathbf{c}) \right) \right) \right].$$

Here, $(\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-) \sim \mathcal{D}$ indicates $\mathbf{c} \sim \mathcal{D}_c, \mathbf{x}_0^+ \sim \pi_\theta(\mathbf{x}_0^+ | \mathbf{c}), \mathbf{x}_0^- \sim \pi_\theta(\mathbf{x}_0^- | \mathbf{c})$.

3.3 DIFFUSION-DPO

The key observation of adapting DPO to diffusion models is that the diffusion denoising objective $\mathcal{L}_{\text{DM}}$ serves as an implicit reward. Specifically, Diffusion-DPO (Wallace et al., 2024) defines a reward for an image x at timestep t as the improvement of the fine-tuned model ϵ_θ over the reference model ϵ_{ref} , specifically $r_\theta(\mathbf{c}, \mathbf{x}, t) \propto \delta_\theta(\mathbf{c}, \mathbf{x}_t, t)$, where

$$\delta_\theta(\mathbf{c}, \mathbf{x}_t, t) := -\left(\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)\|_2^2 - \|\epsilon - \epsilon_{\text{ref}}(\mathbf{x}, \mathbf{c}, t)\|_2^2 \right). \quad (2)$$

The resulting preference objective is

$$\mathcal{L}_{\text{Diff-DPO}}(\theta) = -\mathbb{E}_{(\mathbf{c}, \mathbf{x}^+, \mathbf{x}^-) \sim \mathcal{D}, t} \left[\log \sigma \left(\beta T \omega(\lambda_t) (\delta_\theta(\mathbf{c}, \mathbf{x}^+, t) - \delta_\theta(\mathbf{c}, \mathbf{x}^-, t)) \right) \right],$$

where T is the number of diffusion steps and $\omega(\lambda_t)$ reweights timestep contributions.

4 METHOD

4.1 LISTWISE PREFERENCE OPTIMIZATION FOR DIFFUSION MODELS

Existing preference alignment methods for diffusion models, such as Diffusion-DPO, are built on the Bradley–Terry (BT) model, which assumes pairwise preferences. In this work, we adopt listwise preference optimization, $\mathbf{x}^{(1)} \succ \mathbf{x}^{(2)} \succ \dots \succ \mathbf{x}^{(m)}$, which leverages the ranking as a more fine-grained information for preferences. To capture this higher-order structure, we adopt the *Plackett–Luce* (PL) model (Plackett, 1975), a probabilistic model for listwise rankings.

Reward Under Plackett–Luce Model. Let $\mathcal{G} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\}$ denote a group of m candidate images generated under prompt $\mathbf{c}$, ranked by a user as $\mathbf{x}^{(1)} \succ \mathbf{x}^{(2)} \succ \dots \succ \mathbf{x}^{(m)}$. The PL model defines the probability of this ranking as

$$p_{\text{PL}}(\mathbf{x}^{(1)} \succ \mathbf{x}^{(2)} \succ \dots \succ \mathbf{x}^{(m)} \mid \mathbf{c}) = \prod_{j=1}^m \frac{\exp(r(\mathbf{c}, \mathbf{x}^{(j)}))}{\sum_{k=j}^m \exp(r(\mathbf{c}, \mathbf{x}^{(k)}))},$$

where $r(\mathbf{c}, \mathbf{x})$ denotes the latent reward of the image $\mathbf{x}$ conditioned on the prompt $\mathbf{c}$. This formulation can be viewed as applying a softmax over each suffix sublist $\mathbf{x}^{(j)}, \mathbf{x}^{(j+1)}, \dots, \mathbf{x}^{(m)}$, ensuring that at every step the selected item is assigned a higher preference than all remaining lower-ranked candidates.

Objective for Diffusion-LPO We would like to finetune the diffusion model $p_{\theta}(\mathbf{x}_0 \mid \mathbf{c})$ with the RLHF objective as in Equation 1. For diffusion models, the KL-divergence regularizes the whole diffusion chain, and the reward is defined as $r(\mathbf{c}, \mathbf{x}_0) = \mathbb{E}_{p_{\theta}(\mathbf{x}_{1:T} \mid \mathbf{x}_0, \mathbf{c})} [R(\mathbf{c}, \mathbf{x}_{0:T})]$, where $(\mathbf{x}_1, \dots, \mathbf{x}_T)$ are variables defined on the diffusion path. Maximizing the reward from the objective of Plackett–Luce model, we can get the listwise Diffusion-DPO objective:

$$\begin{aligned} \mathcal{L}_{\text{Diffusion-LPO}}(\theta) = & - \mathbb{E}_{(\mathbf{c}, \mathbf{x}_0^{(1:m)}) \sim \mathcal{D}, \mathbf{x}_{1:T}^{(1:m)} \sim p_{\theta}(\mathbf{x}_{1:T}^{(1:m)} \mid \mathbf{x}_0^{(1:m)})} \\ & \sum_{j=1}^m \left[\beta \log \frac{p_{\theta}(\mathbf{x}_{0:T}^{(j)} \mid \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(j)} \mid \mathbf{c})} - \log \sum_{k=j}^m \exp \left(\beta \log \frac{p_{\theta}(\mathbf{x}_{0:T}^{(k)} \mid \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(k)} \mid \mathbf{c})} \right) \right]. \end{aligned} \quad (3)$$

Let $\epsilon_{\theta}(\cdot, \mathbf{c}, t)$ and $\epsilon_{\text{ref}}(\cdot, \mathbf{c}, t)$ be the learned and reference denoisers. Using the standard Diffusion-DPO surrogate $\Delta_{\theta}(\mathbf{x}_{0:T}^{(j)} \mid \mathbf{c}) \approx T \omega(\lambda_t) \delta_{\theta}(\mathbf{c}, \mathbf{x}_t^{(j)}, t)$, where $\delta_{\theta}(\mathbf{c}, \mathbf{x}_t, t)$ is defined in Equation 2, the listwise objective becomes

$$\begin{aligned} \mathcal{L}_{\text{Diffusion-LPO}}(\theta) = & - \mathbb{E}_{(\mathbf{c}, \mathbf{x}_0^{(1:m)}) \sim \mathcal{D}, \mathbf{x}_t^{(1:m)} \sim p_{\theta}(\mathbf{x}_t^{(1:m)} \mid \mathbf{x}_0^{(1:m)})} \\ & \sum_{j=1}^m \left[\beta T \omega(\lambda_t) \delta_{\theta}(\mathbf{c}, \mathbf{x}_t^{(j)}, t) - \log \sum_{k=j}^m \exp \left(\beta T \omega(\lambda_t) \delta_{\theta}(\mathbf{c}, \mathbf{x}_t^{(k)}, t) \right) \right]. \end{aligned} \quad (4)$$

The proof of deriving Equation (4) from the RLHF objective (Equation 1) can be found in Appendix B.1. As listwise optimization is a more general form of pairwise, it is possible to extend Diffusion-LPO with other pairwise DPO methods. We use DSPO (Cai et al., 2025) as an example and provide the derivation and relevant discussion in Appendix E. We observe improvement when using listwise preferences in DSPO compared with original pairwise DSPO, with detailed results shown in Appendix E.2.

Besides Diffusion-LPO, we also found similar existing works that also aim to optimize at the list level. We include a discussion of the comparison between our method and theirs in Appendix B.2.

Constructing Listwise Groups. The Pick-a-Pic dataset provides user preferences in a pairwise form, e.g., $\mathbf{x}_a \succ \mathbf{x}_b$ and $\mathbf{x}_b \succ \mathbf{x}_c$ under the same prompt $\mathbf{c}$. To construct a more faithful representation of human feedback, we form a list of preferences from the above pairwise preferences as $\mathbf{x}_a \succ \mathbf{x}_b \succ \mathbf{x}_c$. We conduct aggregation as directed acyclic graphs (DAG) of preferences and extract listwise paths, which we treat as ranking sublists. We provide statistics of transforming the pick-a-pic dataset into listwise in Appendix A. An example of the preference group can be found in Figure 5.

4.2 ADVANTAGE OF LISTWISE PREFERENCE MODELING

Given a list of preferences, prior work on pairwise optimization (Wallace et al., 2024) can be extended naively. One can find the optimal policy model by minimizing the objective below, where we refer to it as Group Pairwise DPO (GP-DPO):

$$L_{\text{GP-DPO}} = -\mathbb{E}_{(\mathbf{c}, \mathbf{x}_0^{(1:m)}) \sim \mathcal{D}, \mathbf{x}_{1:T}^{(1:m)} \sim p_{\theta}(\mathbf{x}_{1:T}^{(1:m)} | \mathbf{x}_0^{(1:m)}, \mathbf{c})} \sum_{1 \leq j < k \leq m} \log \sigma \left(r(\mathbf{c}, \mathbf{x}^{(j)}) - r(\mathbf{c}, \mathbf{x}^{(k)}) \right). \quad (5)$$

Such an objective is also derived in Chen et al. (2025). Conducting pairwise optimization in a group of size m will introduce $m(m-1)/2$ pairs for comparison with equal importance. For our proposed method, denote the score $s_{\theta}^{(j)} := \frac{p_{\theta}(\mathbf{x}_{0:T}^{(j)} | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(j)} | \mathbf{c})}$, which is the part of reward $R(\mathbf{c}, \mathbf{x}_{0:T}^{(j)})$ related to policy model up to scalar β . In objective (4), when treating $\mathbf{x}^{(j)}$ as the positive sample and optimizing over its set of negative samples $\{\mathbf{x}^{(j+1)}, \dots, \mathbf{x}^{(m)}\}$, it yields the following reward difference

$$P_{\text{Diffusion-LPO}}(\mathbf{x}^{(j)} > \mathbf{x}^{(k)}, \forall k \in \{j+1, \dots, m\}) \propto \beta \log s_{\theta}^{(j)} - \log \left(\sum_{k=j}^m (s_{\theta}^{(k)})^{\beta} \right). \quad (6)$$

Besides, we can find for Group Pairwise DPO with objective (5), the reward difference when treating $\mathbf{x}^{(j)}$ as the positive sample and computing with all the preference pairs as

$$P_{\text{GP-DPO}}(\mathbf{x}^{(j)} > \mathbf{x}^{(k)}, \forall k \in \{j+1, \dots, m\}) \propto \beta \log s_{\theta}^{(j)} - \frac{1}{m-j+1} \sum_{k=j+1}^m \log((s^{(j)})^{\beta} + (s^{(k)})^{\beta}). \quad (7)$$

The detailed analysis can be found in Appendix B.3. GP-DPO and Diffusion-LPO derive different reward functions in terms of human preference of $\mathbf{x}^{(j)}$ over the negative samples. Compare their rewards for negative samples, we can find

$$\begin{aligned} \frac{1}{m-j+1} \sum_{k=j+1}^m \log((s^{(j)})^{\beta} + (s^{(k)})^{\beta}) &\leq \frac{1}{m-j+1} \sum_{k=j+1}^m \max_{k:k>j} \log((s^{(j)})^{\beta} + (s^{(k)})^{\beta}) \\ &\leq \log \left(\sum_{k=j}^m (s_{\theta}^{(k)})^{\beta} \right). \end{aligned} \quad (8)$$

From the above analysis, we observe that GP-DPO leads to an underestimate of the aggregate reward of negative samples, since it reduces the ranking into equally weighted pairwise comparisons as shown in (7). In particular, the reward assigned to negatives under GP-DPO is upper-bounded by the corresponding term in Diffusion-LPO, as shown in (8). This underestimation inflates the margin between the current positive sample $\mathbf{x}^{(j)}$ and its negatives. In contrast, Diffusion-LPO directly normalizes over the entire negative group as shown in (6), which ensures higher-ranked samples are favored over the lower-ranked ones.

5 EXPERIMENT

5.1 EXPERIMENT SETUP

Models, Datasets, and Evaluations. We finetune two text-to-image diffusion backbones: Stable Diffusion 1.5 (SD1.5) (Rombach et al., 2022) and SDXL (Podell et al., 2023). Training data is drawn from the Pick-a-Pic v1 preference dataset (Kirstain et al., 2023), and our listwise data construction follows the procedure described in Appendix A. We evaluate models on: (1) **General Text-to-Image Alignment:** We use the prompts from three datasets for evaluation: Pick-a-Pic test dataset (Kirstain et al., 2023), Parti-Prompts (Yu et al., 2022), HPSV2 test dataset (Wu et al., 2023). (2) **Image Editing:** We use the instructions on the InstructPix2Pix dataset for evaluation. (3) **Personalized Preference Generation:** Testing whether models adapt to individual user preference embeddings. We train the SD1.5 model with the pipeline in PPD (Dang et al., 2025) and evaluate the personalization performance on the Pick-a-Pic test dataset. Comprehensive details of the evaluation pipeline and prompts are given in Appendix C.1. Implementation details are in Appendix C.2.

Baselines. We consider baselines of: SFT (supervised fine-tuning on preferred images), pairwise Diffusion-DPO (Wallace et al., 2024), and DSPO (Zhu et al., 2025), along with the original pre-trained SD1.5 and SDXL. Besides finetuning on these baselines, we also include Diffusion-DPO*, SD1.5, and SDXL tuned on the Pick-a-Pic v2 preference dataset, which has almost twice the data size as Pick-a-Pic v1. We directly use the released checkpoint for evaluation. Notice that we only consider baselines that do not require additional reward models or evaluators for a fair comparison.

Evaluation. For text-to-image alignment and image editing, we adopt five widely used automatic evaluators: PickScore (Kirstain et al., 2023), HPSV2 (Wu et al., 2023), CLIP score (Radford et al., 2021), Image Reward (Xu et al., 2023), and Aesthetic Score (AES) (Schuhmann, 2022). For personalized preference generation, where no standard automatic metric exists, we follow recent evaluation practice in (Dang et al., 2025) and rely on GPT-4o (Hurst et al., 2024) as a judge, prompting it to rate which generated image better aligns with a user’s historical preferences.

5.2 GENERAL EVALUATION: TEXT-TO-IMAGE ALIGNMENT

Table 1: Winrate results over original SD1.5 and SDXL for Diffusion-LPO compared with other baselines. For abbreviation, we use “PS”, “IM”, and “AES” to represent PickScore, Image Reward, and Aesthetics score. “Diff.” is short for “Diffusion”. **Note: best results trained under Pick-a-Pic v1 (Kirstain et al., 2023) are in boldface. Diffusion-DPO* is trained under Pick-a-Pic v2 dataset, with almost twice the data size than v1.** We directly use the checkpoint trained by (Wallace et al., 2024). All results are averaged over generations from 5 random seeds. ↑: higher values indicate better performance.

Dataset	Method	Stable Diffusion 1.5					Stable Diffusion XL				
		PS ↑	HPS ↑	CLIP↑	IM↑	AES↑	PS↑	HPS ↑	CLIP↑	IM ↑	AES ↑
Pick-a-Pic	SFT	73.4%	80.7%	56.3%	74.9%	71.7%	22.5%	39.3%	46.6%	37.0%	24.3%
	Diff.-DPO*	73.3%	69.8%	57.1%	61.7%	63.4%	77.1%	78.4%	66.0%	70.0%	50.8%
	Diff.-DPO	68.2%	70.7%	54.8%	65.2%	60.1%	73.0%	81.1%	64.5%	66.4%	47.1%
	DSPO	73.1%	82.6%	57.8%	75.2%	70.9%	59.2%	77.4%	56.6%	62.9%	45.9%
	Diff.-LPO	80.4%	74.6%	58.5%	68.1%	64.0%	77.1%	87.1%	64.0%	73.7%	51.5%
Parti-Prompts	SFT	65.0%	79.0%	54.2%	69.7%	72.3%	24.2%	39.6%	47.9%	40.9%	32.0%
	Diff.-DPO*	67.3%	64.9%	53.7%	60.2%	62.3%	70.0%	80.5%	64.6%	74.0%	58.6%
	Diff.-DPO	59.7%	64.0%	52.9%	61.6%	58.9%	66.8%	77.7%	56.0%	68.7%	56.0%
	DSPO	63.8%	78.2%	54.3%	70.3%	71.1%	55.5%	78.1%	54.2%	65.5%	55.4%
	Diff.-LPO	71.9%	69.7%	57.4%	65.9%	62.5%	72.8%	82.9%	60.1%	73.6%	59.7%
HPSV2	SFT	75.1%	85.0%	56.4%	77.7%	72.9%	20.4%	40.1%	47.8%	42.3%	25.9%
	Diff.-DPO*	76.5%	71.0%	57.4%	64.4%	67.2%	72.5%	79.2%	59.3%	70.4%	55.5%
	Diff.-DPO	69.5%	70.9%	51.5%	65.8%	62.0%	72.1%	80.0%	59.2%	70.8%	47.7%
	DSPO	74.4%	85.3%	56.8%	77.4%	71.6%	58.4%	77.3%	47.8%	63.7%	48.2%
	Diff.-LPO	82.9%	76.6%	57.7%	69.0%	65.1%	74.6%	85.0%	56.3%	72.3%	51.9%

We report text-to-image generation results in Table 1. Diffusion-LPO achieves substantial gains over its pairwise baselines and even surpasses Diffusion-DPO trained on Pick-a-Pic v2, on most metrics. On SD1.5, Diffusion-LPO shows clear improvements over pairwise Diffusion-DPO. In particular, the PickScore increases by more than 12% across all evaluation sets, indicating more reliable alignments with human preferences. On SDXL, Diffusion-LPO consistently outperforms all baselines. Notably, Diffusion-LPO improves the PickScore win rate on Parti-prompts with 6% and HPS score win rate on HPSV2 with 5% compared with its pairwise baseline, Diffusion-DPO. While SFT achieves competitive results on SD1.5, its performance drops significantly on SDXL, with win rates falling below 50%. This degradation is likely due to a mismatch between the relatively lower quality of training data and the higher intrinsic quality of SDXL generations. In contrast, Diffusion-LPO exploits relative ranking information effectively, enabling the model to better capture human preferences and outperform its pairwise baseline, Diffusion-DPO.

Figure 3 qualitatively illustrates these improvements. Diffusion-LPO demonstrates a stronger capability to handle fine-grained structures. For the last column, although all methods produce visually appealing images of the dolphin mid-jump, Diffusion-LPO generates dolphin tails with smoother

curvature and more accurate shape, highlighting its superiority in capturing detailed features. More qualitative illustrations of Diffusion-LPO over other baselines are shown in Appendix D.1.



Figure 3: Images generated from original SDXL, Diffusion-DPO, DSPO, and Diffusion-LPO. Diffusion-LPO demonstrates improved image generation quality over other baselines regarding general aesthetics and detail handling. The last column indicates the zoom-in parts.

5.3 IMAGE EDITING

We further evaluate Diffusion-LPO in the context of instruction-based image editing. Following prior work, we adopt the InstructPix2Pix dataset (Brooks et al., 2023), which provides paired source images, natural language edit instructions, and target outputs. We perform editing with Stable Diffusion 1.5, comparing the original model and its tuned variants. For editing, we apply SDEdit (Meng et al., 2021) with a noise strength of 0.6. Diffusion-LPO achieves notable improvements over its pairwise counterpart, Diffusion-DPO, with win-rate gains of 9.1% on PickScore and 4.4% on HPS. These results indicate that listwise supervision provides stronger alignment signals for following editing instructions. Qualitative results are shown in Figure 8 in Appendix D.2.

5.4 PERSONALIZED PREFERENCE ALIGNMENT

Beyond general text-to-image alignment, we further evaluate Diffusion-LPO under the personalized preference setting. Following the pipeline of Dang et al. (2025), we condition the diffusion model on user embeddings extracted by a vision-language model that summarizes each user’s historical preferences (see Appendix C.3 for details).

We use SD1.5 as the backbone and evaluate it on the Pick-a-Pic test set, which is partitioned into held-in users (seen during training) and held-out users (unseen during training). As a baseline, we adopt Personalized Preference Diffusion (PPD) (Dang et al., 2025), which aligns user embeddings with diffusion models using the Diffusion-DPO objective. We then replace the pairwise loss with our listwise formulation, denoted as PPD-LPO. Incorporating Diffusion-LPO into the personaliza-

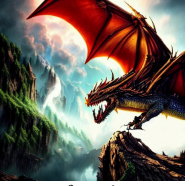
User Profile	SD 1.5	SD 1.5+Profile	PPD	PPD-LPO
This user prefers images that ... capture the beauty of nature and technology coexisting , such as trains traveling through scenic landscapes , evoke feelings of adventure and progress. ...				
<i>cinematic still of highly reflective stainless steel train in the desert, at sunset</i>				
The preferred images share common traits such as high visual quality, detailed textures... and immersive scenes . They often feature majestic dragons in natural landscapes , evoking a sense of wonder and fantasy. ...				
<i>a majestic vicious dragon habitat landscape, boss fight scene, refractions realistic, cinematic lighting, antview, natural, horrific atmosphere, sharp focus</i>				
The user prefers images that have high visual quality, with detailed and dynamic elements, vibrant colors... The user values depth, lighting, and color saturation in their preferred images, ...				
<i>swirling water tornados epic fantasy</i>				
The preferred images consistently exhibit more detailed character design , richer colors, and engaging backgrounds... dispreferred images lack these qualities, appearing flatter, less detailed...				
<i>paladin in full plate armor, 3d render, realistic, photorealism</i>				

Figure 4: Images generated under Diffusion-LPO with personal preference alignment with other baselines. User profiles are summarized by VLM. “SD 1.5+Profile” represents images generated using SD 1.5 with user profile appended to the caption. We highlight the user preferences in **green** and dispreferences in **red**.

tion pipeline yields consistent gains: the win rate improves from 71.1% to 72.3% on held-in users and from 70.3% to 80.2% on held-out users, demonstrating that listwise optimization both enhances personalization for known users and substantially improves generalization to unseen users. Qualitative analysis is shown in Figure 4. While the original PPD improves alignment relative to the untuned SD1.5 model, PPD-LPO further enhances both visual appeal and fine-grained detail.

5.5 ABLATION: RANK ENFORCEMENT FROM DIFFUSION-LPO

To further assess the effectiveness of Diffusion-LPO, we compare it against Group-Pairwise DPO (GP-DPO), with the objective in Equation 5 to finetune Stable Diffusion 1.5. Both objectives are trained under the same objective. For each prompt, we calculate the scores of images generated by

win rates across all evaluation sets, indicating stronger preference alignment. This result demonstrates that, by enforcing the diffusion model to optimize on human preference in ranking structure, Diffusion-LPO achieves better alignment quality, validating our theoretical analysis in Section 4.2

Table 2: Winrate for Diffusion-LPO on SD1.5 in comparison to SD1.5 trained under GP-DPO.

Dataset	PS↑	HPS↑	CLIP↑	IM↑	AES↑
Pick-a-Pic	52.3%	51.3%	50.6%	50.8%	53.7%
HPSv2	52.4%	54.0%	50.6%	51.3%	49.1%
Parti-Prompts	52.4%	53.0%	50.8%	51.4%	52.1%

6 CONCLUSION

In this work, we introduce Diffusion-LPO, a novel framework for aligning text-to-image models using listwise preference optimization. By modeling full preference rankings with the Plackett-Luce model, our approach surpasses pairwise methods like Direct Preference Optimization (DPO). Experiments on Stable Diffusion 1.5 and SDXL demonstrate that Diffusion-LPO produces outputs with superior coherence and human preference alignment, highlighting the efficacy of listwise supervision for fine-tuning generative models.

REFERENCES

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislaw Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in neural information processing systems*, 35:38176–38189, 2022.
- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.
- Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.
- Miaomiao Cai, Simiao Li, Wei Li, Xudong Huang, Hanting Chen, Jie Hu, and Yunhe Wang. Dspo: Direct semantic preference optimization for real-world image super-resolution. *arXiv preprint arXiv:2504.15176*, 2025.
- Lichang Chen, Chen Zhu, Jiu hai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoenybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. In *International Conference on Machine Learning*, pp. 7935–7952. PMLR, 2024.
- Renjie Chen, Wenfeng Lin, Yichen Zhang, Jiangchuan Wei, Boyuan Liu, Chao Feng, Jiao Ran, and Mingyu Guo. Towards self-improvement of diffusion models via group preference optimization. *arXiv preprint arXiv:2505.11070*, 2025.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.
- Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8020–8030, 2025.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.
- Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.
- Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.
- Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

- Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024a.
- Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2024b.
- Zijing Hu, Fengda Zhang, and Kun Kuang. D-fusion: Direct preference optimization for aligning diffusion models with visually consistent samples. *arXiv preprint arXiv:2505.22002*, 2025.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Daniel R Jiang, Alex Nikulkov, Yu-Chia Chen, Yang Bai, and Zheqing Zhu. Improving generative ad text on facebook using reinforcement learning. *arXiv preprint arXiv:2507.21983*, 2025.
- Shyamgopal Karthik, Huseyin Coskun, Zeynep Akata, Sergey Tulyakov, Jian Ren, and Anil Kag. Scalable ranked preference optimization for text-to-image generation. *arXiv preprint arXiv:2410.18013*, 2024.
- Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.
- Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.
- Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18465–18475, 2025.
- Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *Advances in Neural Information Processing Systems*, 37:24897–24925, 2024.
- Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(5):7, 2024.
- Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2404–2420, 2025.
- Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. Inpo: Inversion preference optimization with reparametrized ddim for efficient diffusion model alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28629–28639, 2025.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

- Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PmLR, 2016.
- Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- Christoph Schuhmann. Laion-aesthetics. <https://laion.ai/blog/laion-aesthetics/>, 2022. Accessed: 2023 - 11- 10.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18990–18998, 2024.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.
- Haoyuan Sun, Bo Xia, Yongzhe Chang, and Xueqian Wang. Generalizing alignment paradigm of text-to-image generation with preferences through f-divergence minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27644–27652, 2025a.
- Haoyuan Sun, Bo Xia, Yifei Zhao, Yongzhe Chang, and Xueqian Wang. Positive enhanced preference alignment for text-to-image models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

- Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024.
- Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.
- Fu-Yun Wang, Yunhao Shui, Jintan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 1313–1322, 2018.
- Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023.
- Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densdpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.
- Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrp: Unleashing grp on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.
- Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiabin Chen, Weihang Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.
- Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.
- Xin Yu, Zelin He, Ying Sun, Lingzhou Xue, and Runze Li. Understanding the statistical accuracy-communication trade-off in personalized federated learning with minimax guarantees. In *Forty-second International Conference on Machine Learning*, 2025.
- Daoan Zhang, Guangchen Lan, Dong-Jun Han, Wenlin Yao, Xiaoman Pan, Hongming Zhang, Mingxiao Li, Pengcheng Chen, Yu Dong, Christopher Brinton, et al. Seppo: Semi-policy preference optimization for diffusion alignment. *arXiv preprint arXiv:2410.05255*, 2024.
- Lingfan Zhang, Chen Liu, Chengming Xu, Kai Hu, Donghao Luo, Chengjie Wang, Yanwei Fu, and Yuan Yao. When preferences diverge: Aligning diffusion models with minority-aware adaptive dpo. *arXiv preprint arXiv:2503.16921*, 2025.
- Yulai Zhao, Masatoshi Uehara, Gabriele Scalia, Sunyuan Kung, Tommaso Biancalani, Sergey Levine, and Ehsan Hajiramezanali. Adding conditional control to diffusion models with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A PICK-A-PIC DATASET LISTWISE CONSTRUCTION

We restructure the Pick-a-Pic dataset into user-prompt specific directed acyclic graphs (DAGs) of image rankings. Specifically, we collect all images generated under the same prompt and annotated by the same user, and construct a graph where each image corresponds to a node. Whenever the user provides a pairwise annotation indicating that image $\mathbf{x}_A$ is preferred over image $\mathbf{x}_B$, we add a directed edge $\mathbf{x}_A \rightarrow \mathbf{x}_B$. By aggregating all such edges, we obtain a DAG that encodes the transitive structure of the user’s annotations. From these DAGs, we can extract consistent listwise rankings of images (e.g. $\mathbf{x}_A \succ \mathbf{x}_B \succ \mathbf{x}_C$) whenever sufficient pairwise information is available, thus transforming independent pairwise judgments into richer listwise preference orders.

Table 3: Statistics of Pick-a-Pic after listwise group construction.

List Size	# of list
2	225,656
3	81,704
4	42,113
5	22,216
6	14,191
7	8,953
≥ 8	23,087

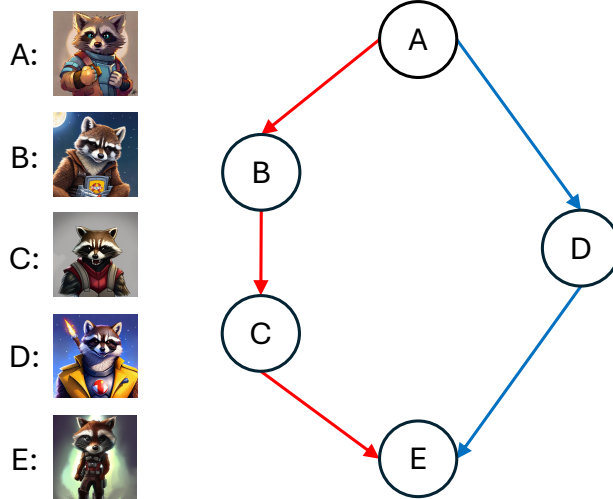


Figure 5: An example of a group with preferences that forms a DAG. The arrow pointing from image $\mathbf{x}_A$ to image $\mathbf{x}_B$ represents human preference: $\mathbf{x}_A \succ \mathbf{x}_B$. The prompt is “Rocket Raccoon, furry art, fanart, digital painting”. Here, valid rank list will be $(\mathbf{x}_A \succ \mathbf{x}_B \succ \mathbf{x}_C \succ \mathbf{x}_E)$ and $(\mathbf{x}_A \succ \mathbf{x}_D \succ \mathbf{x}_E)$.

Data usability. Table 3 summarizes the statistics of the reformulated lists. In total, we analyze 511,840 preference pairs. Among them, 225,656 pairs (44.09%) form lists of size 2, which naturally reduce to standard pairwise DPO training. Importantly, 274,895 pairs (53.71%) belong to lists of size > 2 , making them directly applicable to listwise preference optimization. Only 11,289 pairs (2.20%) fall into inconsistent groups and are discarded. Such inconsistencies correspond to cyclic preferences (e.g., $\mathbf{x}_A \succ \mathbf{x}_B$, $\mathbf{x}_B \succ \mathbf{x}_C$, $\mathbf{x}_C \succ \mathbf{x}_A$), which cannot be represented as DAGs and thus reflect low-quality or noisy annotations. The very small fraction of such cases demonstrates that the overwhelming majority of human annotations are internally consistent and can be reliably aggregated into rankings. In our training, lists of size 2 still contribute through Diffusion-DPO, while longer lists provide richer preference information for Diffusion-LPO.

Constructing lists from DAGs. Each group is represented as a directed acyclic graph (DAG), where nodes correspond to images and edges encode user-indicated pairwise preferences. To transform this structure into training examples for listwise optimization, we enumerate valid topological paths within each DAG. Each path yields an ordered list of images $\{x^{(1)} \succ x^{(2)} \succ \dots \succ x^{(m)}\}$ that is consistent with all local pairwise edges. These lists are then used directly in the Plackett–Luce likelihood of our objective. In practice, multiple valid paths may exist within a single DAG; we use all valid paths for Diffusion-LPO.

B DIFFUSION-LPO OBJECTIVE

B.1 DIFFUSION-LPO OBJECTIVE

Recall the RLHF objective:

$$\max_{p_\theta} \mathbb{E}_{\mathbf{c}, \mathbf{x}_0} [r(\mathbf{c}, \mathbf{x}_0)] - \beta \mathbb{D}_{\text{KL}} [p_\theta(\mathbf{x}_0|\mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_0|\mathbf{c})], \quad (9)$$

where β is a parameter that controls how much $p_\theta(\mathbf{x}_0|\mathbf{c})$ deviates from $p_{\text{ref}}(\mathbf{x}_0|\mathbf{c})$.

We define $R(\mathbf{c}, \mathbf{x}_{0:T})$ as the reward on the whole diffusion chain to define the reward $r(\mathbf{c}, \mathbf{x}_0)$ as $r(\mathbf{c}, \mathbf{x}_0) = \mathbb{E}_{p(\mathbf{x}_{1:T}|\mathbf{x}_0, \mathbf{c})} [R(\mathbf{c}, \mathbf{x}_{0:T})]$. Given 9, we have

$$\begin{aligned} & \min_{p_\theta} -\mathbb{E}_{p_\theta(\mathbf{x}_0|\mathbf{c})} \left[\frac{1}{\beta} r(\mathbf{c}, \mathbf{x}_0) \right] + \mathbb{D}_{\text{KL}} [p_\theta(\mathbf{x}_0|\mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_0|\mathbf{c})] \\ & \leq \min_{p_\theta} -\mathbb{E}_{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})} \left[\frac{1}{\beta} R(\mathbf{c}, \mathbf{x}_{0:T}) \right] + \mathbb{D}_{\text{KL}} [p_\theta(\mathbf{x}_{0:T}|\mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})] \\ & = \min_{p_\theta} \mathbb{E}_{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})} \left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})} - \frac{1}{\beta} R(\mathbf{c}, \mathbf{x}_{0:T}) \right] \\ & = \min_{p_\theta} \mathbb{E}_{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})} \left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c}) \exp(R(\mathbf{c}, \mathbf{x}_{0:T})/\beta)} \right] \\ & = \min_{p_\theta} \mathbb{E}_{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})} \left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{\frac{1}{Z(\mathbf{c})} p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c}) \exp(R(\mathbf{c}, \mathbf{x}_{0:T})/\beta)} - \log Z(\mathbf{c}) \right] \\ & = \min_{p_\theta} \mathbb{D}_{\text{KL}} [p_\theta(\mathbf{x}_{0:T}|\mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c}) \exp(R(\mathbf{c}, \mathbf{x}_{0:T})/\beta) / Z(\mathbf{c})]. \end{aligned} \quad (10)$$

where $Z(\mathbf{c}) = \sum_{\mathbf{x}} p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c}) \exp(r(\mathbf{c}, \mathbf{x}_0)/\beta)$ is the partition function, and we get rid of $\log Z(\mathbf{c})$ as it is independent from p_θ . The optimal $p_\theta^*(\mathbf{x}_{0:T}|\mathbf{c})$ of Equation (10) has a unique closed-form solution:

$$p_\theta^*(\mathbf{x}_{0:T}|\mathbf{c}) = p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c}) \exp(R(\mathbf{c}, \mathbf{x}_{0:T})/\beta) / Z(\mathbf{c}),$$

Therefore, we have the reparameterization of the reward function

$$R(\mathbf{c}, \mathbf{x}_{0:T}) = \beta \log \frac{p_\theta^*(\mathbf{x}_{0:T}|\mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})} + \beta \log Z(\mathbf{c}).$$

Plugging this into the definition of r , we have the following

$$r(\mathbf{c}, \mathbf{x}_0) = \beta \mathbb{E}_{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0, \mathbf{c})} \left[\log \frac{p_\theta^*(\mathbf{x}_{0:T}|\mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})} \right] + \beta \log Z(\mathbf{c}).$$

Substituting this reward reparameterization into the maximum likelihood objective of the Plackett–Luce model, the partition function cancels out, and we get a maximum likelihood objective defined on diffusion models, for the group $(\mathbf{c}, \mathbf{x}_0^{(1)}, \dots, \mathbf{x}_0^{(m)})$:

$$\mathcal{L}_{\text{Diffusion-LPO}}(\theta) = \log \prod_{j=1}^m \frac{\exp \left(\beta \mathbb{E}_{p_\theta(\mathbf{x}_{1:T}^{(j)}|\mathbf{x}_0^{(j)}, \mathbf{c})} \left[\log \frac{p_\theta^*(\mathbf{x}_{0:T}^{(j)}|\mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(j)}|\mathbf{c})} \right] \right)}{\sum_{k=j}^m \exp \left(\beta \mathbb{E}_{p_\theta(\mathbf{x}_{1:T}^{(k)}|\mathbf{x}_0^{(k)}, \mathbf{c})} \left[\log \frac{p_\theta^*(\mathbf{x}_{0:T}^{(k)}|\mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(k)}|\mathbf{c})} \right] \right)}.$$

By applying Jensen's inequality, we can push the expectation out, and with some simplification, we can have,

$$\begin{aligned} \mathcal{L}_{\text{Diffusion-LPO}}(\theta) = & -\mathbb{E}_{(\mathbf{c}, \mathbf{x}_0^{(1:m)}) \sim \mathcal{D}, \mathbf{x}_{1:T}^{(1:m)} \sim p_\theta(\mathbf{x}_{1:T}^{(1:m)} | \mathbf{x}_0^{(1:m)})} \\ & \sum_{j=1}^m \left[\beta \log \frac{p_\theta(\mathbf{x}_{0:T}^{(j)} | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(j)} | \mathbf{c})} - \log \sum_{k=j}^m \exp \left(\beta \log \frac{p_\theta(\mathbf{x}_{0:T}^{(k)} | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(k)} | \mathbf{c})} \right) \right]. \end{aligned}$$

Since sampling from $p_\theta(\mathbf{x}_{1:T} | \mathbf{x}_0, \mathbf{c})$ is intractable, we utilize the forward process $q(\mathbf{x}_{1:T} | \mathbf{x}_0)$ for approximation. As $\mathbf{x}_{1:T}^{(j)} \sim q(\mathbf{x}_{1:T} | \mathbf{x}_0^{(j)}, \mathbf{c})$ for $j = 1, \dots, m$,

$$\begin{aligned} L_{\text{approx}}(\theta) = & -\sum_{j=1}^m \left[\beta \mathbb{E}_{\mathbf{x}_{1:T}^{(j)}} \left[\log \frac{p_\theta(\mathbf{x}_{0:T}^{(j)} | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(j)} | \mathbf{c})} \right] - \log \sum_{k=j}^m \exp \left(\beta \mathbb{E}_{\mathbf{x}_{1:T}^{(k)}} \left[\log \frac{p_\theta(\mathbf{x}_{0:T}^{(k)} | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(k)} | \mathbf{c})} \right] \right) \right] \\ = & -\sum_{j=1}^m \left[\beta \mathbb{E}_{\mathbf{x}_{1:T}^{(j)}} \left[\sum_{t=1}^T \log \frac{p_\theta(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})} \right] \right. \\ & \left. - \log \sum_{k=j}^m \exp \left(\beta \mathbb{E}_{\mathbf{x}_{1:T}^{(k)}} \left[\sum_{t=1}^T \log \frac{p_\theta(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})} \right] \right) \right] \\ = & -\sum_{j=1}^m \left[\beta \mathbb{E}_{\mathbf{x}_{1:T}^{(j)}} \left[T \mathbb{E}_t \log \frac{p_\theta(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})} \right] \right. \\ & \left. - \log \sum_{k=j}^m \exp \left(\beta \mathbb{E}_{\mathbf{x}_{1:T}^{(k)}} \left[T \mathbb{E}_t \log \frac{p_\theta(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})} \right] \right) \right] \\ = & -\sum_{j=1}^m \left[\beta T \mathbb{E}_{t, \mathbf{x}_{t-1}^{(j)}, \mathbf{x}_t^{(j)}} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})} \right] \right. \\ & \left. - \log \sum_{k=j}^m \exp \left(\beta T \mathbb{E}_{t, \mathbf{x}_{t-1}^{(k)}, \mathbf{x}_t^{(k)}} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})} \right] \right) \right] \\ = & -\sum_{j=1}^m \left[\beta T \mathbb{E}_{t, \mathbf{x}_t^{(1:m)}, \mathbf{x}_{t-1}^{(j)}} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})} \right] \right. \\ & \left. - \log \sum_{k=j}^m \exp \left(\beta T \mathbb{E}_{t, \mathbf{x}_t^{(1:m)}, \mathbf{x}_{t-1}^{(k)}} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})} \right] \right) \right], \end{aligned} \tag{11}$$

where $\mathbf{x}_{1:T}^{(j)} \sim q(\mathbf{x}_{1:T} | \mathbf{x}_0^{(j)}, \mathbf{c})$, $\mathbf{x}_t^{(j)} \sim q(\mathbf{x}_t | \mathbf{x}_0^{(j)})$ and $\mathbf{x}_{t-1}^{(j)} \sim q(\mathbf{x}_{t-1} | \mathbf{x}_t^{(j)}, \mathbf{x}_0^{(j)})$ for $j = 1, \dots, m$. Since the stagewise negative Plackett–Luce term $-\sum_j [s_j - \log \sum_{k \geq j} \exp(s_k)]$ is convex in the score vector s , by Jensen's inequality we can push $\mathbb{E}_{t, \mathbf{x}_t^{(1:m)}}$ outside and obtain an upper bound:

$$\begin{aligned} L_{\text{approx}}(\theta) \leq & -\mathbb{E}_{t, \mathbf{x}_t^{(1:m)}} \sum_{j=1}^m \left[\beta T \mathbb{E}_{\mathbf{x}_{t-1}^{(j)}} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(j)} | \mathbf{x}_t^{(j)}, \mathbf{c})} \right] \right. \\ & \left. - \log \sum_{k=j}^m \exp \left(\beta T \mathbb{E}_{\mathbf{x}_{t-1}^{(k)}} \left[\log \frac{p_\theta(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{t-1}^{(k)} | \mathbf{x}_t^{(k)}, \mathbf{c})} \right] \right) \right]. \end{aligned}$$

Using the standard reverse-Gaussian parameterization of the diffusion model, each inner expectation reduces to a KL-difference that, in turn, yields the denoising-error form with the usual timestep

weight $\omega(\lambda_t)$:

$$L_{\text{approx}}(\theta) = -\mathbb{E}_{\mathbf{c}, \mathbf{x}_0^{(1:m)}, t} \sum_{j=1}^m \left[\beta T \omega(\lambda_t) \left(-\|\epsilon^{(j)} - \epsilon_\theta(\mathbf{x}_t^{(j)}, \mathbf{c}, t)\|_2^2 + \|\epsilon^{(j)} - \epsilon_{\text{ref}}(\mathbf{x}_t^{(j)}, \mathbf{c}, t)\|_2^2 \right) \right. \\ \left. - \log \sum_{k=j}^m \exp \left(\beta T \omega(\lambda_t) \left(-\|\epsilon^{(k)} - \epsilon_\theta(\mathbf{x}_t^{(k)}, \mathbf{c}, t)\|_2^2 + \|\epsilon^{(k)} - \epsilon_{\text{ref}}(\mathbf{x}_t^{(k)}, \mathbf{c}, t)\|_2^2 \right) \right) \right],$$

where $\epsilon^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_t^{(j)} = \sqrt{\alpha_t} \mathbf{x}_0^{(j)} + \sqrt{1 - \alpha_t} \epsilon^{(j)}$, and $\lambda_t = \alpha_t^2 / \sigma_t^2$.

B.2 COMPARISON WITH EXISTING RANK PREFERENCE OPTIMIZATION METHODS.

We found similar existing works that also aim to optimize at the list level. Defining the score as $\mathbf{s}^{(j)} = \|\epsilon^{(j)} - \epsilon_\theta(\mathbf{x}_t^{(j)}, \mathbf{c}, t)\|_2^2 - \|\epsilon^{(j)} - \epsilon_{\text{ref}}(\mathbf{x}_t^{(j)}, \mathbf{c}, t)\|_2^2$, Chen et al. (2025) derives a group pairwise loss (GPO) in the form of $L_{\text{GPO}} = \sum_{j=1}^m [(m - 2j + 1) \mathbf{s}^{(j)}]$ from the pairwise DPO loss, and introduces a standard rewards $\mathcal{A}_j = \frac{\mathbf{r} - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$ for $\mathbf{r} = \{r^{(j)}\}_{j=1}^m$ representing the reward for each image. The reward is scored by some external evaluators. The image reward replaces the term $(m - 2j + 1)$, and the final objective becomes,

$$L_{\text{GPO}} = \sum_{j=1}^m [\mathcal{A}_j (\|\epsilon^{(j)} - \epsilon_\theta(\mathbf{x}_t^{(j)}, \mathbf{c}, t)\|_2^2 - \|\epsilon^{(j)} - \epsilon_{\text{ref}}(\mathbf{x}_t^{(j)}, \mathbf{c}, t)\|_2^2)].$$

Also, Karthik et al. (2024) proposes RankDPO, which uses lambda loss (Wang et al., 2018) for the list of images:

$$L_{\text{RankDPO}} = -\sum_{j=1}^m \sum_{k=j}^m \Delta_{j,k} \log \sigma(-\beta(\mathbf{s}^{(j)} - \mathbf{s}^{(k)})),$$

where $\Delta_{j,k} = |(2^{\phi(j)} - 1) - (2^{\phi(k)} - 1)| \cdot \left| \frac{1}{\log(1+\tau(j))} - \frac{1}{\log(1+\tau(k))} \right|$ represents the weight between image pairs $(\mathbf{x}^{(j)}, \mathbf{x}^{(k)})$, with $\phi(j), \phi(k)$ and $\tau(j), \tau(k)$ being the true scores and true ranks of $(\mathbf{x}^{(j)}, \mathbf{x}^{(k)})$. As the true scores are not available, Karthik et al. (2024) considers evaluation models such as HPS, Pick Score, Image Reward, etc., to compute the simulated scores for the images.

Different from existing methods, Diffusion-LPO does not require any additional scoring model to learn the ranked preference, as (a) the evaluator can cause extra computation cost, and (b) the automated responses from evaluators cannot replace the human annotation. As the goal is to learn user preferences, evaluator feedback may cause misleading information and contradict the actual human feedback.

B.3 DETAIL COMPARIOSN BETWEEN GP-DPO AND DUFFUSION-LPO

Let's consider the data point $\mathbf{x}^{(j)}$ in the preference ranking and its negative samples $\{\mathbf{x}^{(j+1)}, \dots, \mathbf{x}^{(m)}\}$. When using the Group Pairwise DPO, the preference of data point $\mathbf{x}^{(k)}$ over its negative samples is reflected in the pairwise comparison:

$$\log \left(P_{\text{GP-DPO}}(\mathbf{x}^{(j)} > \mathbf{x}^{(k)}, \forall k \in \{j+1, \dots, m\}) \right) = \log \left(\prod_{k=j+1}^m \sigma(r(\mathbf{c}, \mathbf{x}^{(j)}) - r(\mathbf{c}, \mathbf{x}^{(k)})) \right).$$

For simplicity, we omit the expectation over c and $\mathbf{x}_{0:T}^{(1:m)}$ in our analysis. With the notation we stated before, $s_\theta^{(i)} := \frac{p_\theta(\mathbf{x}_{0:T}^{(i)} | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^{(i)} | \mathbf{c})}$, we can find

$$\sum_{k=j+1}^m \log \sigma \left(\beta \log s^{(j)} - \beta \log s^{(k)} \right) = \sum_{k=j+1}^m \left[\beta \log s^{(j)} - \log \left((s^{(j)})^\beta + (s^{(k)})^\beta \right) \right] \\ \propto \beta \log s^{(j)} - \frac{1}{m-j+1} \sum_{k=j+1}^m \log \left((s^{(j)})^\beta + (s^{(k)})^\beta \right).$$

Here, by scaling it with the scalar $1/(m-j+1)$, we can easily check the effect of pairwise preference optimization over a sub-list $\{\mathbf{x}^{(j)}, \dots, \mathbf{x}^{(m)}\}$ in our setting.

For Diffusion-LPO in (3), we can find that the advantage of $\mathbf{x}^{(j)}$ over its negative samples is clearly characterized by a softmax function

$$\begin{aligned} \log \left(P_{\text{Diffusion-LPO}}(\mathbf{x}^{(j)} > \mathbf{x}^{(k)}, \forall k \in \{j+1, \dots, m\}) \right) &= \beta \log s^{(j)} - \log \sum_{k=j}^m \exp \left(\beta \log s^{(k)} \right) \\ &= \beta \log s^{(j)} - \log \sum_{k=j}^m (s^{(k)})^\beta. \end{aligned}$$

Up to now, we have explicitly characterized how GP-DPO and Diffusion-LPO optimize the sub-list preference $\{\mathbf{x}^{(j)}, \dots, \mathbf{x}^{(m)}\}$.

C EXPERIMENT SETTINGS

C.1 EVALUATION DATA DETAILS

We provide the dataset details we used as follows:

Pick-a-Pic dataset : Pick-a-Pic is a large-scale, publicly available dataset of human preferences for text-to-image generation (Kirstain et al., 2023). It is collected via a web application where users submit prompts, receive images from multiple diffusion backbones, and indicate their preferences between pairs of images. Version 1 (v1) contains over 500,000 examples, while the updated Version 2 (v2) extends this to more than one million. Each entry includes a prompt, two generated images, and a label for the preferred image (or a tie). In our experiments, as the Pick-a-Pic v2 is currently not available online, we use Pick-a-Pic v1 for training and its held-out test set for evaluation.

Parti-Prompts : The Parti-Prompts dataset (Yu et al., 2022) was introduced to evaluate compositional and semantic capabilities of large-scale text-to-image models. It contains carefully curated prompts spanning diverse categories, such as objects, attributes, styles, and complex multi-object relations. The prompts are designed to systematically probe generation fidelity, compositionality, and visual grounding, making it a widely used benchmark for assessing the generalization of diffusion-based models. We adopt Parti-Prompts for standardized evaluation of text-to-image generation quality.

HPSV2 : The Human Preference Score v2 (HPS v2) dataset (Wu et al., 2023) is built from over 25,000 prompts and 98,000 images generated by Stable Diffusion, accompanied by 25,205 human preference annotations collected from the Stable Foundation Discord community. Each annotation compares multiple candidate images for the same prompt, with the user selecting a preferred image. This dataset was used to train a preference classifier fine-tuned from CLIP, which defines the HPS metric—a strong predictor of human judgments. We use the HPS v2 test set to benchmark how well our aligned diffusion models capture human preferences.

InstructPix2Pix : InstructPix2Pix (Brooks et al., 2023) is a large-scale dataset for instruction-based image editing. It was synthetically generated by combining GPT-3 for producing editing instructions and Stable Diffusion with Prompt-to-Prompt for creating paired before/after images. The resulting dataset comprises over 450,000 examples of input images, natural language editing instructions, and corresponding edited outputs. Models trained on this dataset can generalize to real-world user-written instructions at inference time. In our evaluation, we use a standardized subset of 1,000 test samples to assess performance on instruction-guided image editing.

C.2 IMPLEMENTATION DETAILS

Hyperparameters. We adopt AdamW for training SD1.5 and Adafactor for training SDXL. The learning rate is set to 1×10^{-8} with linear warmup, scaled by effective batch size. The global

batch size is 2048: for pairwise methods, the effective batch contains 2048 pairs, and for listwise methods, it contains 2048 groups. For SD1.5, we set $\beta = 2000$ for Diffusion-DPO, Diffusion-LPO, and $\beta = 0.001$ for DSPO and DSPO-LPO. For SDXL, we set $\beta = 5000$ for Diffusion-DPO, Diffusion-LPO, and $\beta = 3000$ for DSPO and DSPO-LPO, following previous work.

C.3 PERSONALIZED PREFERENCE IMAGE GENERATION

We introduce the pipeline of our personalized preference alignment tasks here. The setting is adapted from Dang et al. (2025).

Personalized Preference Alignment PPD is a framework designed to align diffusion models with individual-level human preferences. Personalization would lead to better generalization performance across different individuals (Yu et al., 2025). The pipeline operates in two stages. In the first stage, a vision-language model (VLM) summarizes each user’s historical pairwise preference data into a dense user embedding. This embedding captures stylistic and semantic signals specific to the user, derived from only a few examples. In the second stage, the diffusion model is fine-tuned with these user embeddings as additional conditioning, injected via cross-attention layers. Training uses a variant of the Diffusion-DPO objective, which optimizes the model to generate images aligned with each user’s preferences while maintaining regularization against a reference model.

User Embedding Generation Each user is represented using a small set of 4-shot preference pairs (caption, preferred image, dispreferred image). These examples are processed by the multimodal VLM LLaVA-OneVision (with a Qwen2 language backbone and multi-image capability). The final user embedding is obtained by extracting the hidden state of the last token of this profile from the Qwen2 encoder. A Chain-of-Thought (CoT) prompting is used: the VLM describes and compares each preference pair, then generates a textual user profile. The template of prompting can be found in 5.

D ADDITIONAL QUALITATIVE ANALYSIS

D.1 TEXT TO IMAGE QUALITATIVE RESULTS

We show more examples of images generated by Diffusion-LPO. Figure 6 provides examples for Diffusion-LPO over other baselines for SD1.5. From the top to the bottom, the prompts are:

- Beautiful girl;
- Woman Argentina;
- Realistic owl;
- mechanical bee flying in nature, electronics, motors, wires, buttons, lcd;
- beautiful, tranquil garden of love and peace;
- JEM doll 80s in a fur coat in the snow, thick outlines, bright colors, digital art, hard edges, detailed, anime style, dynamic pose, character design, art by sora kim, rinotuna, ilya kuvshinov;
- Duck.

Figure 7 shows more examples for Diffusion-LPO over other baselines for SDXL. From the top to the bottom, the prompts are:

- A blue airplane in a blue, cloudless sky;
- Portrait of an anime maid by Krenz Cushart, Alphonse Mucha, and Ilya Kuvshinov;
- Two Somali friends sitting and watching a Studio Ghibli movie;
- A jellyfish sleeping in a space station pod;
- A painting by Raffaello Sanzi portraying Kajol and symbiots Riot during the Renaissance era, showcased on Artstation;
- A person holding a very small slice on pizza between their fingers.



Figure 6: Qualitative illustrations for Diffusion-LPO over other baselines for SD1.5.



Figure 7: Qualitative illustrations for Diffusion-LPO over other baselines for SDXL.

D.2 IMAGE EDIT QUALITATIVE RESULTS

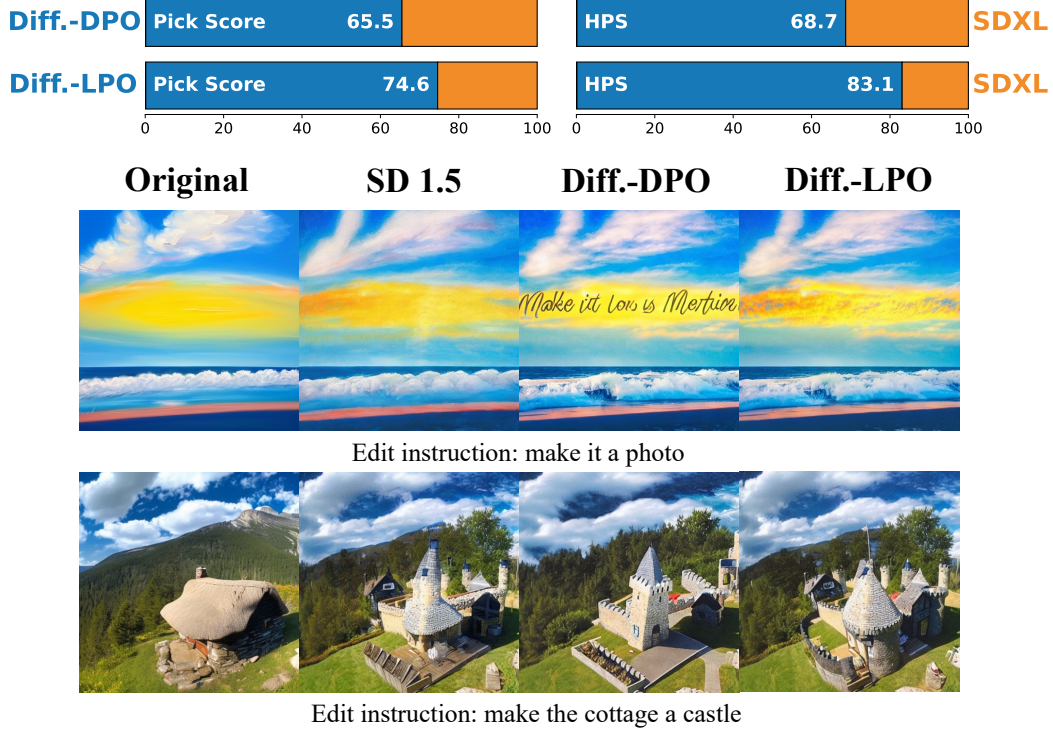


Figure 8: We evaluate the Pick Score and HPS for the generated images, where Diffusion-LPO shows consistent improvement. For qualitative results, Diffusion-LPO faithfully follows the edit instructions while producing high-quality pictures. Diffusion-LPO produces more faithful and higher-quality edits. In the first example, only Diffusion-LPO successfully transforms a paint-like image into a realistic photograph, accurately reflecting the edit instruction. In the second example, while all methods modify the cottage into a castle as instructed, Diffusion-LPO generates a visually sharper and more coherent castle with realistic architectural details, underscoring its superior ability to handle fine-grained edits. More results are available in Figure 9.

We show more qualitative results for image editing in Figure 9. Edit prompts and original images are from the InstructPix2Pix dataset.

E GENERALIZATION FROM OTHER DPO FRAMEWORKS

The idea of switching from pairwise to listwise can be applied to most of the families of Diffusion DPO works. Here, we show one example of adapting Listwise preference into a DPO work, DSPO (Zhu et al., 2025). DSPO leverages score matching from human preference, and its original objective is,

$$\min_{\theta} \omega(t) \|\nabla_{\mathbf{x}_t} \log p_{\theta}(\mathbf{x}_t | \mathbf{c}) - (\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{c}) + \gamma \nabla_{\mathbf{x}_t} \log p_{\theta}(\mathbf{y} | \mathbf{x}_t, \mathbf{c}))\|_2^2, \quad (12)$$

where $\omega(t)$ is the time-dependent function for score matching Song et al. (2021). DSPO model human preference with $p(\mathbf{y} | \mathbf{x}_t, \mathbf{c}) = p(\mathbf{x}_t \succ \mathbf{x}_t^l | \mathbf{x}_t^l, \mathbf{c}) = \sigma(r(\mathbf{c}, \mathbf{x}_t) - r(\mathbf{c}, \mathbf{x}_t^l))$ and reward calculated by $r(\mathbf{c}, \mathbf{x}_t) \propto -(\|\epsilon_{\theta}(\mathbf{x}_{t+1}, t+1, \mathbf{c}) - \epsilon_{t+1}\|_2^2 - \|\epsilon_{\text{ref}}(\mathbf{x}_{t+1}, t+1, \mathbf{c}) - \epsilon_{t+1}\|_2^2)$. Plug into the objective(Equation 12), their loss function becomes,

$$\mathcal{L}_{\text{DSPO}} = A(t) \|\epsilon_{\theta, t+1} - \epsilon_{t+1} - \lambda \gamma (1 - \sigma(r(\mathbf{c}, \mathbf{x}_t) - r(\mathbf{c}, \mathbf{x}_t^l)))(\epsilon_{\theta, t+1} - \epsilon_{\text{ref}, t+1})\|. \quad (13)$$

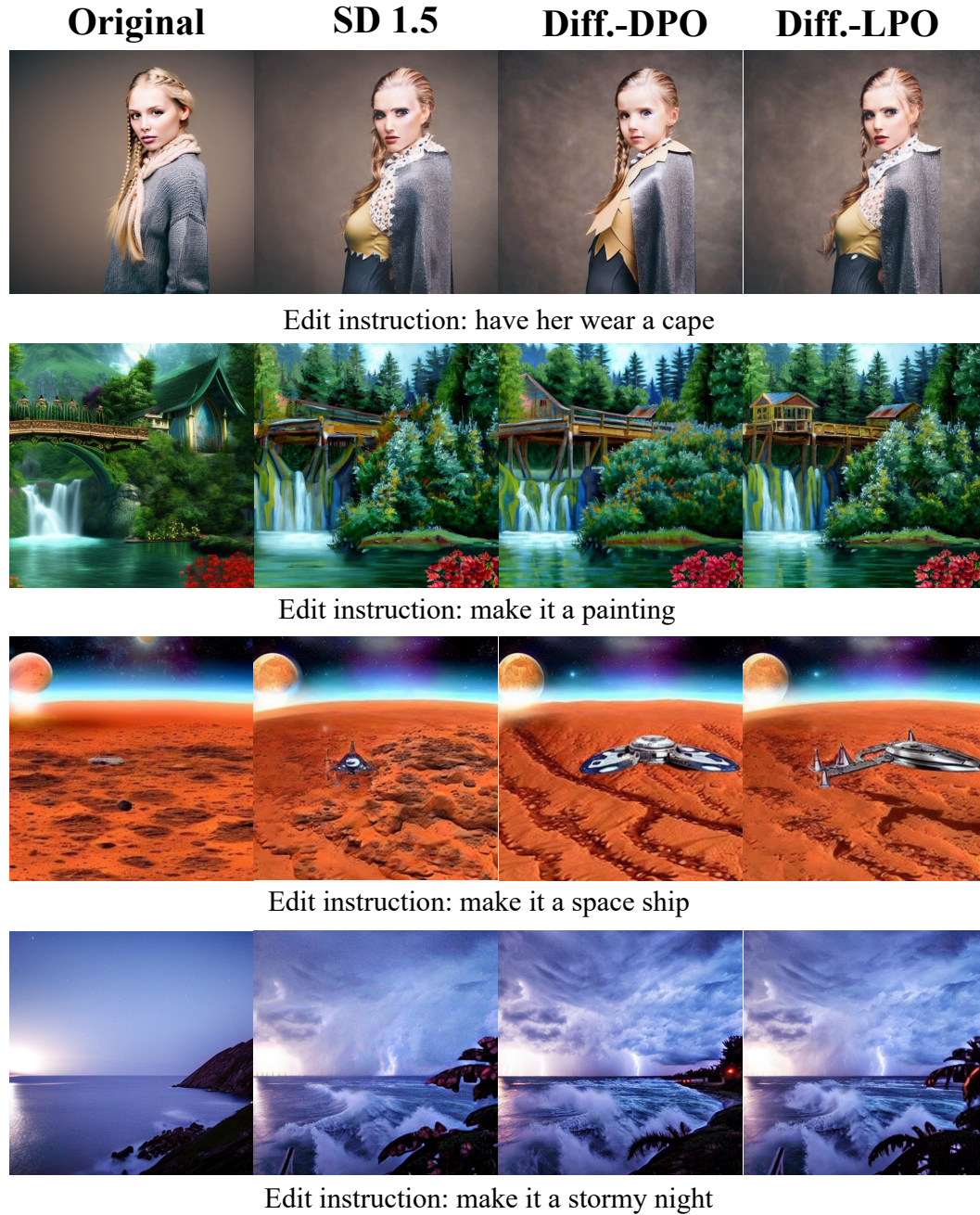


Figure 9: Qualitative results for image edit using SD1.5, finetuned over Diffusion-LPO and other baselines.

To extend DSPO into listwise preference modeling, given a list of images with preferences $(\mathbf{x}^{(1)} \succ \mathbf{x}^{(2)} \succ \dots \succ \mathbf{x}^{(m)})$ we can score-match the human preference through

$$p(\mathbf{y}|\mathbf{x}_t, \mathbf{c}) = p(\mathbf{x}_t^{(i)} \succ \mathbf{x}_t^{(i+1)} \succ \dots \succ \mathbf{x}_t^{(m)} | \mathbf{x}_t^{(i+1)} \succ \dots \succ \mathbf{x}_t^{(m)}, \mathbf{c}) = \frac{\exp(r(c, x^{(i)}))}{\sum_{k=i}^m \exp(r(c, x^{(k)}))}. \quad (14)$$

Therefore, a listwise preference optimization for DSPO (we denoted as DSPO-LPO) is,

$$\mathcal{L}_{\text{DSPO-LPO}} = A(t) \|\epsilon_{\theta, t+1} - \epsilon_{t+1} - \lambda \gamma \left(1 - \frac{\exp(r(c, x^{(i)}))}{\sum_{k=i}^m \exp(r(c, x^{(k)}))}\right) (\epsilon_{\theta, t+1} - \epsilon_{\text{ref}, t+1})\|. \quad (15)$$

In practice, we choose the highest rank image, $\mathbf{x}^{(1)}$, as the target of score matching ($i = 1$). A detailed derivation can be found at Appendix E.1.

E.1 DSPO-LPO OBJECTIVE

We start with the original objective of Equation 12. By modeling $p_\theta(\mathbf{y}|\mathbf{x}_t, \mathbf{c})$ as $\frac{\exp(r(\mathbf{c}, \mathbf{x}^{(i)}))}{\sum_{k=i}^m \exp(r(\mathbf{c}, \mathbf{x}^{(k)}))}$, we convert Equation 12 into:

$$\min_{\theta} \omega(t) \|\nabla_{\mathbf{x}_t} \log \frac{p_\theta(\mathbf{x}_t|\mathbf{c})}{p(\mathbf{x}_t|\mathbf{c})} - \gamma \nabla_{\mathbf{x}_t} \log \frac{\exp(r(\mathbf{c}, \mathbf{x}^{(i)}))}{\sum_{k=i}^m \exp(r(\mathbf{c}, \mathbf{x}^{(k)}))}\|_2^2 \quad (16)$$

$$= \min_{\theta} \omega(t) \|\nabla_{\mathbf{x}_t} \log \frac{p_\theta(\mathbf{x}_t|\mathbf{c})}{p(\mathbf{x}_t|\mathbf{c})} - \gamma \left(1 - \frac{\exp(r(\mathbf{c}, \mathbf{x}^{(i)}))}{\sum_{k=i}^m \exp(r(\mathbf{c}, \mathbf{x}^{(k)}))}\right) \nabla_{\mathbf{x}_t} r(\mathbf{c}, \mathbf{x}_t)\|_2^2 \quad (17)$$

From the derivation in (Zhu et al., 2025), we can write the first term as:

$$\nabla_{\mathbf{x}_t} \log \frac{p_\theta(\mathbf{x}_t|\mathbf{c})}{p(\mathbf{x}_t|\mathbf{c})} \approx \epsilon_{\theta, t+1} - \epsilon_{t+1}, \quad (18)$$

and that $\nabla_{\mathbf{x}_t} r(\mathbf{c}, \mathbf{x}_t)$ can be expressed in terms of the reference model:

$$\nabla_{\mathbf{x}_t} r(\mathbf{c}, \mathbf{x}_t) \approx \epsilon_{\theta, t+1} - \epsilon_{\text{ref}, t+1}. \quad (19)$$

Thus the objective simplifies to:

$$\mathcal{L}_{\text{DSPO}} = A(t) \|\epsilon_{\theta, t+1} - \epsilon_{t+1} - \lambda \gamma \left(1 - \sigma(r(\mathbf{c}, \mathbf{x}_t) - r(\mathbf{c}, \mathbf{x}^{(j)}))\right) (\epsilon_{\theta, t+1} - \epsilon_{\text{ref}, t+1})\|_2^2, \quad (26)$$

where $A(t)$ absorbs the timestep-dependent weight and λ is a constant factor arising from the Gaussian parameterization.

E.2 TEXT-TO-IMAGE ALIGNMENT OF DSPO-LPO

We show the quantitative result of DSPO and DSPO-LPO in Table 4. We can observe that, by incorporating listwise preference optimization with DSPO, we further boost the performance across all metrics.

Table 4: Winrate results over original SDXL for Diffusion-LPO compared with DSPO, both fine-tuned on SDXL.

Dataset	Method	PS $\uparrow$	HPS $\uparrow$	CLIP $\uparrow$	IM $\uparrow$	AES $\uparrow$
Pick-a-Pic	DSPO	59.2%	77.4%	56.6%	62.9%	45.9%
	DSPO+LPO	77.1%	87.1%	64.0%	73.7%	51.5%
Parti-Prompts	DSPO	55.5%	78.1%	54.2%	65.5%	55.4%
	DSPO+LPO	72.8%	82.9%	60.1%	73.1%	59.7%
HPSV2	DSPO	58.4%	77.3%	47.8%	63.7%	48.2%
	DSPO+LPO	74.6%	85.0%	56.3%	72.3%	51.9%

System Prompt
You are an expert in image aesthetics and have been asked to predict which image a user would prefer based on the examples provided.
COT Assistant Prompt
You will be shown a few examples of preferred and dispreferred images that a user has labeled. Here is Pair 1: Here is the caption: [Caption for Pair 1] Here is Image 1: [Image 1] Here is Image 2: [Image 2] Prediction of user preference: [1 or 2] [...] Here is Pair 4: Here is the caption: [Caption for Pair 1] Here is Image 1: [Image 1] Here is Image 2: [Image 2] Prediction of user preference: [1 or 2] 1. Describe each image in terms of style, visual quality, and image aesthetics. 2. Explain the differences between the two images in terms of style, visual quality, and image aesthetics. 3. After you have described all of the images, summarize the differences between the preferred and dispreferred images into a user profile. Format your response as follows for the four pairs of images: Pair 1: Image 1: [Description] Image 2: [Description] Differences: [Description] [...] Pair 4: Image 1: [Description] Image 2: [Description] Differences: [Description] User Profile: [Description]
Additional Assistant Prompt for User Preference Prediction
Finally, you are provided with a new pair of images, unlabeled by the user. Your task is to predict which image the user would prefer based on the previous examples you have seen. Format your response as follows: Prediction of user preference: [1 or 2]

Table 5: Instructions for Embedding Generation and User Preference Prediction

F THE USE OF LARGE LANGUAGE MODELS

The paper is primarily written by the authors, with only minor assistance from large language models (LLMs). In particular, LLMs were used for light editing tasks such as polishing grammar, refining phrasing, and suggesting alternative wordings. All technical content, derivations, and experimental design were conceived and carried out by the authors. No LLMs were used for generating research ideas, implementing methods, or conducting experiments.

G ETHICS STATEMENTS

We build on publicly available datasets and models, following established practices in prior work (Wallace et al., 2024; Podell et al., 2023; Rombach et al., 2022; Kirstain et al., 2023). All models are obtained from official repositories such as HuggingFace, and we respect associated licenses. While the Pick-a-Pic dataset may contain potentially unsafe content (e.g., depictions of violence or nudity), our use is strictly limited to research purposes within the dataset’s intended scope. No additional human annotation was conducted in this study.

We are mindful of the broader societal impacts of generative models. Our work is intended to advance alignment techniques so that diffusion models better capture user preferences in a transparent and responsible manner. We do not release user-identifiable data and take care to avoid privacy risks. We also emphasize that preference alignment should be deployed in ways that respect diversity, minimize potential harm, and promote accessibility. All experiments were conducted in accordance with ethical and legal standards, with a commitment to openness, reproducibility, and integrity in reporting results.

H REPRODUCIBILITY STATEMENT

We use 8 H100-80G for our experiments, with GPU hours around 50 hours for finetuning SDXL for each experiment. We will release the model checkpoint and code source in the future.

I LIMITATION

In this work, the maximum list size is restricted to 8. We believe further exploration with larger lists and a systematic study of how list size influences optimization can provide additional insights. Moreover, evaluating Diffusion-LPO on a broader range of preference datasets would help further verify its effectiveness and generality.