Margin-Aware Preference Optimization for Aligning Diffusion Models Without Reference

Jiwoo Hong[1*], **Sayak Paul**[2*], **Noah Lee**[1], **Kashif Rasul**[2], **James Thorne**[3], **Jongheon Jeong**[4]

[1]KAIST AI
[2]Hugging Face
[3]Theia Insights
[4]Korea University

Abstract

Modern preference alignment methods, such as DPO, rely on divergence regularization to a reference model for training stability—but this creates a fundamental problem we call "reference mismatch." In this paper, we investigate the negative impacts of reference mismatch in aligning text-to-image (T2I) diffusion models, showing that larger reference mismatch hinders effective adaptation given the same amount of data, *e.g.,* as when learning new artistic styles, or personalizing to specific objects. We demonstrate this phenomenon across text-to-image (T2I) diffusion models and introduce **margin-aware preference optimization (MaPO)**, a *reference-agnostic* approach that breaks free from this constraint. By directly optimizing the likelihood margin between preferred and dispreferred outputs under the Bradley-Terry model without anchoring to a reference, MaPO transforms diverse T2I tasks into unified pairwise preference optimization. We validate MaPO's versatility across **five** challenging domains: (1) safe generation, (2) style adaptation, (3) cultural representation, (4) personalization, and (5) general preference alignment. Our results reveal that MaPO's advantage grows dramatically with reference mismatch severity, outperforming both DPO and specialized methods like DreamBooth while reducing training time by 15%. MaPO thus emerges as a versatile and memory-efficient method for generic T2I adaptation tasks.

Code — https://github.com/mapo-t2i/mapo
Project — https://mapo-t2i.github.io
Extended version — https://arxiv.org/abs/2406.06424

Introduction

Diffusion models have become dominant for modeling high-dimensional data distributions due to their scalability (Ho, Jain, and Abbeel 2020; Kingma et al. 2021; Rombach et al. 2022; Podell et al. 2024; Peebles and Xie 2023; Esser et al. 2024), handling diverse conditioning modalities including text (Li et al. 2022; Strudel et al. 2022), images (Ho, Jain, and Abbeel 2020; Podell et al. 2024), and audio (Kong et al. 2021; Evans et al. 2024). Their capabilities in human-centered applications have motivated *fine-tuning* for preference alignment in areas like safe generation (Shen et al. 2024; Schramowski et al. 2023), stylistic rendering (Hertz et al. 2024), and personalization (Ruiz et al. 2023; von Rütte et al. 2023).

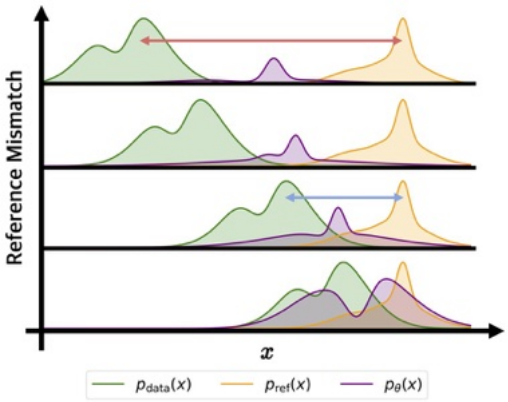


Figure 1: *Reference mismatch*, heterogeneity between the training model's distribution p_θ and the data distribution p_{data}, leads to suboptimal preference learning. MaPO is a reference-free preference optimization method, more robust to such heterogeneity while adapting to the preference.

T2I diffusion model alignment generates outputs reflecting desired attributes through preference optimization (Lee, Liu et al. 2023; Yoon et al. 2023; Fan et al. 2023; Wallace et al. 2023; Li et al. 2024b; Yuan et al. 2024), with reinforcement learning methods treating denoising as multi-step decision-making via proximal policy optimization (Schulman, Wolski et al. 2017). These methods employ reference models with divergence regularization to stabilize training, prevent overfitting, and preserve core capabilities (Ziegler et al. 2020; Wang et al. 2024a; Skalse et al. 2022; Pang et al. 2023).

However, constraining models to specific references limits flexibility in learning new content (Tajwar et al. 2024). We term this *reference mismatch* (Figure 1)—when reference model features differ from preference data, often triggered by using stronger proprietary models for dataset curation (Wu et al. 2023; Lambert et al. 2025). In T2I models, this manifests as stylistic preferences ("cartoon") or distributional biases from limited personalization data (Bianchi et al. 2023;

*These authors contributed equally.

Liu et al. 2024a; Lu et al. 2023; Hertz et al. 2024), affecting all task-specific methods aiming to induce new attributes while preserving general capabilities (Ruiz et al. 2023; Lee et al. 2024a). Addressing these discrepancies is crucial for applying preference alignment to diverse downstream tasks.

We investigate how reference mismatch hinders T2I diffusion model alignment when using direct alignment methods (Wallace et al. 2023), finding adverse effects particularly significant with large distributional gaps. We introduce margin-aware preference optimization (MaPO), a novel reference-agnostic method that defines the Bradley-Terry model score function (Bradley and Terry 1952) directly from training model likelihood and incorporates DDPM loss (Ho, Jain, and Abbeel 2020) to incrementally align data and model distributions. While reference-free alignment has been studied in language modeling (Xu, Sharaf et al. 2024; Hong, Lee, and Thorne 2024; Meng, Xia, and Chen 2024; Gupta et al. 2025), we develop the first such objective for T2I diffusion models.

Our experiments on five representative T2I tasks—safe generation, style learning, cultural representation, personalization, and preference alignment—show that MaPO overcomes reference mismatch challenges while maintaining alignment benefits. For example, MaPO outperforms three preference alignment methods, including InPO (Lu et al. 2025a) and two personalization methods, including DreamBooth (Ruiz et al. 2023), while reducing training time by 15% compared to Diffusion-DPO (Wallace et al. 2023). This study provides a unified framework for T2I tasks and preference learning, empirically validated across five distinct tasks.

Preliminaries

Text-to-image diffusion models Text-to-image (T2I) diffusion models (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022) learn to denoise random noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ into a data sample $x_0 \sim p_{\text{data}}(x_0)$ conditioned on text prompt c. They model a discrete Markov process $p_\theta(x_{t-1}|x_t, c)$ that predicts x_{t-1} from x_t for timesteps $t = T, \ldots, 1$, where x_t follows the forward diffusion process:

$$x_t \sim q(x_t|x_0) \quad \text{where} \quad q(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 \mathbf{I}), \quad (1)$$

with noise schedule parameters α_t and σ_t (Ho, Jain, and Abbeel 2020). The backward denoising process is defined as:

$$p_\theta(x_{0:T}|c) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c). \quad (2)$$

To maximize the likelihood of observed data x_0 under model $p_\theta(x_0|c)$, the evidence lower bound is minimized. Ho, Jain, and Abbeel (2020) parameterized p_θ as a noise predictor $\epsilon_\theta(x_t, c, t)$, yielding an MSE objective with random noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

$$\begin{aligned} L_{\text{DDPM}} &\leq \mathbb{E}_{x_T}\left[-\log p_\theta(x_0 \mid c)\right] \\ &\leq T \cdot \mathbb{E}_{x_0,\epsilon,t}\left[\omega(\lambda_t)\left\|\epsilon - \epsilon_\theta\left(x_t, c, t\right)\right\|^2\right], \end{aligned} \quad (3)$$

where $\omega(\lambda_t)$ depends on the signal-to-noise ratio $\lambda_t = \log(\alpha_t^2/\sigma_t^2)$ (Song and Ermon 2019; Kingma et al. 2021). In practice, a simplified loss is used:

$$\mathcal{L}_{\text{MSE}}(c, x_0) := \mathbb{E}_{\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, c, t\right)\right\|^2\right]. \quad (4)$$

Preference optimization for alignment Alignment fine-tunes generative models to produce human-preferred outputs (Ouyang, Wu et al. 2022). Human preferences are often collected as pairs (x^w, x^l) given prompt c, where x^w ("chosen") is preferred over x^l ("rejected"). The Bradley-Terry model (Bradley and Terry 1952) models the preference as:

$$p(x^w \succ x^l \mid c) = \frac{\exp(r(x^w, c))}{\exp(r(x^w, c)) + \exp(r(x^l, c))}, \quad (5)$$

where $r(x, c)$ denotes the reward function. This approach, popularized in language model alignment (Ziegler et al. 2020; Rafailov et al. 2023), is often combined with reinforcement learning like PPO (Schulman, Wolski et al. 2017) in RLHF:

$$\max_\theta \mathbb{E}_{x\sim p_\theta(x|c)}\left[r(x, c)\right] - \beta \mathbb{D}_{\text{KL}}\left(p_\theta(x|c) \parallel p_{\text{ref}}(x|c)\right), \quad (6)$$

where p_{ref} is the reference model (typically the pre-trained initialization) and β weights the KL constraint. The optimal policy for objective (6) is:

$$p^*(x \mid c) = \frac{1}{Z(c)} \cdot p_{\text{ref}}(x \mid c) \cdot \exp\left(\frac{1}{\beta} \cdot r(x, c)\right), \quad (7)$$

where $Z(c)$ is the partition function. Direct alignment algorithms (DAAs) like DPO (Rafailov et al. 2023) achieve this without RL by directly optimizing the implicit reward. For T2I diffusion models, Wallace et al. (2023) adapted DPO to preferences over diffusion paths $x_{1:T}^w$ and $x_{1:T}^l$:

$$\begin{aligned} &\mathcal{L}_{\text{Diff-DPO}}(c, x_{1:T}^w, x_{1:T}^l) \\ &\quad := -\log\sigma\left(\beta \log \frac{p_\theta(x_{1:T}^w|c)}{p_{\text{ref}}(x_{1:T}^w|c)} - \beta \log \frac{p_\theta(x_{1:T}^l|c)}{p_{\text{ref}}(x_{1:T}^l|c)}\right). \end{aligned} \quad (8)$$

Margin-aware Preference Optimization

In this section, we first establish the concept of *reference mismatch* when aligning text-to-image (T2I) diffusion models and their negative impacts. Then, we propose *margin-aware preference optimization* (MaPO), a novel preference alignment method for diffusion models that aims to mitigate the issue by eliminating the need for a reference model.

Motivation: Reference mismatch problem

We define *reference mismatch* as the divergence (*e.g.,* KL divergence) between the preference data distribution p_{data} and the initial reference model p_{ref}. The negative impacts of reference mismatch have been empirically observed in language models, particularly in DPO training (Guo et al. 2024; Tajwar et al. 2024; Xu et al. 2024; Tang et al. 2024). This issue mainly arises from the key assumption in DPO, namely, that the chosen and rejected samples (x^w, x^l) are drawn from the optimal policy p^* (7) (Rafailov et al. 2023). However, in practice, preference data rarely originate from the optimal policy (Xu et al. 2024; Tang et al. 2024; Liu et al. 2024b), often being generated from external sources (Wallace et al. 2023; Li et al. 2024b; Zhu, Xiao, and Honavar 2025). This discrepancy violates this assumption and hinders optimal policy learning in DPO, highlighting the necessity of addressing the reference mismatch. A possible workaround to address the reference mismatch of DPO is lowering the

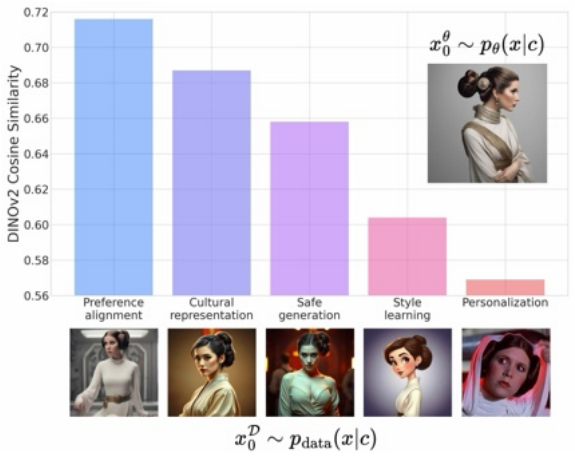


Figure 2: Reference mismatch between the model generation x_0^θ and data $x_0^\mathcal{D}$ quantified by the cosine distance in the DINOv2 embeddings. The *personalization* task has the lowest similarity, implying the highest mismatch.

hyperparameter β (8) to reduce the dependency of p_θ to p_{ref}; however, this approach often triggers performance degradation in generation quality, due to that lowering β also weakens the log-likelihood objective of $p_\theta(x|c)$ (Rafailov et al. 2024; Pal et al. 2024; Shi et al. 2024; Liu, Liu, and Cohan 2024). Therefore, lowering β does not mitigate reference mismatch and its negative impacts but deteriorates the model, making this scenario's necessity of p_{ref} questionable.

Reference mismatch in T2I tasks Similarly, in T2I diffusion models, the optimality of Diffusion-DPO is prone to reference mismatch. As an instance, we quantify the reference mismatch in five representative downstream tasks in T2I diffusion models: general preference alignment (Wallace et al. 2023; Li et al. 2024b), cultural representation (Bianchi et al. 2023; Liu et al. 2024a), safe generation (Schramowski et al. 2023; Kim et al. 2023), style learning (Lu et al. 2023; Hertz et al. 2024), and personalization (Ruiz et al. 2023; Lee et al. 2024a). We measure the reference mismatch with image similarity score using DINOv2 (Oquab et al. 2024) between $x_0^\theta \sim p_\theta(x|c)$ and $(x_0^\mathcal{D}, c) \sim p_{\text{data}}(x|c)$: *i.e.,* less reference mismatch with higher score. In Figure 2, generic preference alignment and personalization tasks were shown to have the smallest and largest reference mismatch out of five tasks. This shows that the degree of reference mismatch significantly varies by task, limiting the versatility of direct alignment methods with reference models like Diffusion-DPO in the downstream tasks of T2I diffusion models.

Approach: Reference-free diffusion alignment

We propose a new preference optimization algorithm that eliminates the need for a reference model in diffusion alignment. Overall, the key idea is to define the reference-agnostic score function in the Bradley-Terry (BT) model.

Objective function of MaPO Given a preference dataset p_{data} of triplets of the form (c, x_0^l, x_0^w), each of which consists of a prompt c and a preference image pair (x_0^w, x_0^l) given c. MaPO optimizes a T2I diffusion model p_θ with:

$$\begin{aligned}&\mathcal{L}_{\text{MaPO}}(c, x_0^w, x_0^l)\\ &\quad := \mathcal{L}_{\text{MSE}}(c, x_0^w) + \tfrac{1}{\beta}\mathcal{L}_{\text{Margin}}(c, x_0^l, x_0^w)\end{aligned} \tag{9}$$

$$\begin{aligned}&\mathcal{L}_{\text{Margin}}(c, x_0^w, x_0^l)\\ &:= -\log\sigma\left(\phi_\beta(\mathcal{L}_{\text{MSE}}(c, x_0^w)) - \phi_\beta(\mathcal{L}_{\text{MSE}}(c, x_0^l))\right),\end{aligned} \tag{10}$$

and $\mathcal{L}_{\text{MSE}}$ is the standard DDPM objective in (3) that maximizes the likelihood for "chosen" pairs (c, x_0^w) in (9), and $\mathcal{L}_{\text{Margin}}$ (10) is the proposed margin-aware regularization that defines the score function in the BT model using the gap of $\mathcal{L}_{\text{MSE}}$ between x_0^w and x_0^l, modulated by a *link function* ϕ_β:

$$\phi_\beta(\ell) := \left(\frac{\ell}{\exp(\ell) - 1}\right)^\beta. \tag{11}$$

In a nutshell, (10) aims to regularize p_θ to ***(i)*** ensure that x^w and x^l achieve sufficient likelihood margin, and ***(ii)*** fuse the term once they have the margin. In this way, MaPO incorporates preference pairs (x^l, x^w) upon simple distribution matching and defines a new preference optimization, which notably requires no reference model.

Joint matching and alignment Supervised fine-tuning (SFT) is one of straightforward approaches for matching the distribution of p_θ to p_{data} (Kumar et al. 2022; Sun 2024). We incorporate the standard diffusion loss (4), computed with the "chosen" samples x^w, into MaPO (9) as an SFT to incrementally match the distribution of p_θ to p_{data} throughout the alignment. While SFT has been conventionally adopted to initially match p_θ before preference learning (Bai et al. 2022; Rafailov et al. 2023; Meng, Xia, and Chen 2024), making overall training multi-stage, this often induces an additional distribution mismatch during the preference learning phase due to static (*i.e.,* off-policy) preference data (Guo et al. 2024). Thus, we adopt SFT within the preference learning stage, consistently matching p_θ to p_{data} while learning the preference to prevent additional mismatches.

Preference learning as a margin regularization We aim to eliminate the use of p_{ref} for preference optimization given the negative impacts of the noisy divergence penalty discussed above. Recall that under the Bradley-Terry model, a preference distribution can be modeled as follows:

$$p(x_1 \succ x_2|c) = \sigma\left(f\left(c, x_1\right) - f\left(c, x_2\right)\right), \tag{12}$$

where $f(c, x)$ represents the general representation of score function that assigns a scalar score to the prompt c and the image x pair. DPO parameterizes f with p_θ and p_{ref} as r_{DPO},

$$r_{\text{DPO}}(x, c) = \beta\log\frac{p_\theta(x, c)}{p_{\text{ref}}(x, c)} + \log Z(c), \tag{13}$$

as $Z(c)$ as a partition function for the prompt c from the maximum entropy reinforcement learning (Wallace et al. 2023; Rafailov et al. 2024). However, misguiding of p_{ref} is one factor that hinders desired preference learning as discussed previously. Furthermore, as implicit reward r_{DPO} is not bounded either way, it is prone to overfitting by $r_{\text{DPO}}(c, x_l)$ and $r_{\text{DPO}}(c, x_w)$ easily diverging to maximize their margin

with logistic loss (12) (Azar et al. 2023; Kim et al. 2024) and eventually deteriorating the model in extreme cases (Liu, Liu, and Cohan 2024; Shi et al. 2024).

From this vein, we introduce bounded link function (11) that can define the score function f in (12) without p_{ref}. Along with the reference-agnostic design, it prevents the excessive divergence problem of r_{DPO} by being bounded within $(0, 1)$. Here, the hyperparameter β of (11) controls the temperature of the score function, allowing (10) to be minimized with less likelihood margin between (c, x_0^w) and (c, x_0^l) when β gets larger. Finally, we weight (10) with β^{-1} to cancel out the proportional impact of β in $\nabla_\theta \mathcal{L}_{\text{Margin}}$, since the gradient of (10) is proportional to β (see Supplementary).

Theoretical justification for reference-free link functions The effectiveness of reference-free methods like MaPO, ORPO (Hong, Lee, and Thorne 2024), and SimPO (Meng, Xia, and Chen 2024) can be understood through the lens of the Bradley-Terry (BT) model's flexibility. The BT model (12) only requires a score function $f(c, x)$ that assigns scalar values preserving preference relationships—it does not mandate any specific functional form. While DPO's log-ratio formulation (13) emerges from maximum entropy RL theory, it represents just one possible instantiation.

The key requirements for a valid score function in preference learning are: (i) monotonicity with respect to generation quality, (ii) bounded outputs to prevent optimization instabilities, and (iii) preservation of preference orderings (Sun, Shen, and Ton 2025; Gupta et al. 2025). MaPO's link function ϕ_β (11) satisfies all these criteria using only the model's likelihood (via $\mathcal{L}_{\text{MSE}}$), without requiring p_{ref}. The bounded nature of $\phi_\beta \in (0, 1)$ particularly addresses the divergence issues of unbounded rewards (Azar et al. 2023; Kim et al. 2024).

This theoretical understanding explains why diverse reference-free formulations succeed, *i.e.,* ORPO's log-odds and SimPO's log-probability are alternative score functions that maintain preference relationships through model-intrinsic measures. By eliminating the divergence penalty, these methods avoid the pitfalls of reference mismatch while still effectively propagating preference signals. Our work extends this principle to diffusion models, demonstrating its applicability across generative modeling paradigms.

Unifying T2I fine-tuning as preference alignment Despite its broad formulation, it has been conventionally believed that applying preference optimization to diverse T2I fine-tuning tasks beyond general preference alignment, *e.g.,* for style adaptation, is limited in practice; this is possibly due to the fact that *reference mismatch* in typical T2I fine-tuning can be more severe than in language alignment. By circumventing the reference mismatch through a *reference-free* alignment, MaPO expands the range of T2I diffusion model fine-tuning tasks where pairwise preference optimization can be effectively applied. Once we have a specific target image x_0 to stipulate as *chosen* image x_0^w and corresponding prompt c, the sampled generation $x_0^l \sim p_\theta(x|c)$ from the T2I diffusion model to be trained can be *rejected* image x_0^l. Thereby, MaPO can be a versatile alignment method that could be generally used for the T2I fine-tuning tasks based on target datasets of the form $(x_0, c) \sim p_{\text{data}}$.

Experiments

We validate the effectiveness and general applicability of MaPO across diverse text-to-image (T2I) diffusion model fine-tuning tasks with ***five*** baselines, three from preference alignment and two for personalization. Specifically, we construct a benchmark of ***five*** representative T2I downstream adaptation scenarios, each with varying degrees of reference mismatch, including the standard preference alignment task:

1. **Safe generation** (Schramowski et al. 2023)
2. **Style learning** (Lu et al. 2023; Hertz et al. 2024)
3. **Cultural representation** (Bianchi et al. 2023)
4. **Personalization** (Ruiz et al. 2023; Lee et al. 2024a)
5. **Preference alignment** (Wallace et al. 2023)

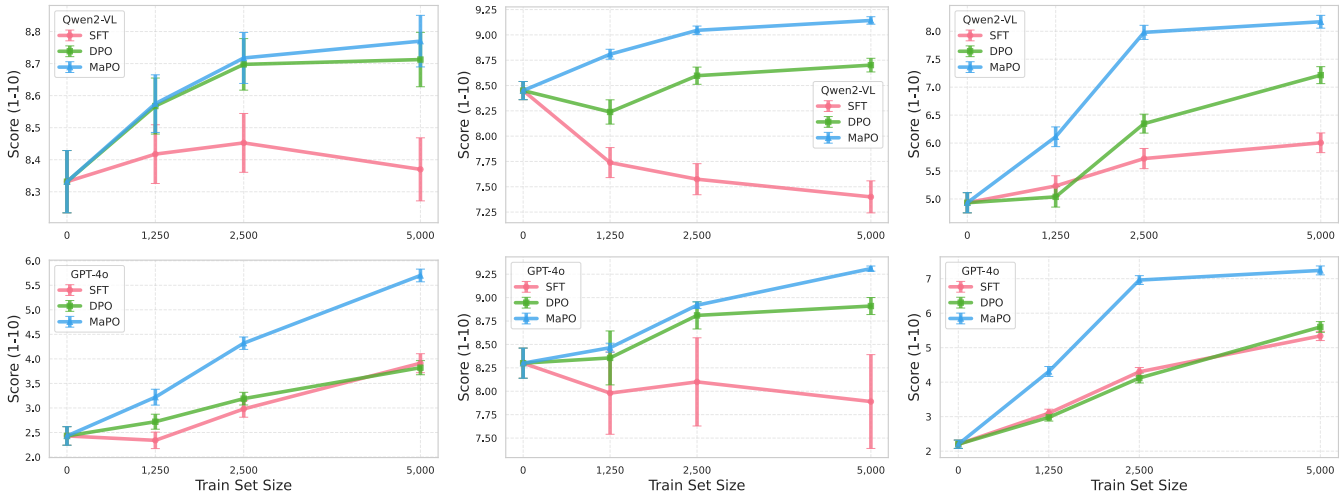
Experimental details

We compare MaPO and other methods by fine-tuning Stable-Diffusion XL (Podell et al. 2024, SDXL), and evaluate them with task-specific metrics, including HPSv2.1 (Wu et al. 2023), and DINOv2 (Oquab et al. 2024), and VLM-as-a-Judge (Chen et al. 2024b,a; Lee et al. 2024b; Yasunaga, Zettlemoyer, and Ghazvininejad 2025). We state the detailed training configurations in the Supplementary.

Specialized preference alignment For a controlled comparison across the tasks under this category, we develop synthetic preference data on top of Pick-a-Pic v2. We sample 20,000 prompts from Pick-a-Pic v2 and extract the core contexts using GPT-3.5-Turbo.[1] Then, we employ FLUX.1-Schnell (Black Forest Labs 2024) to generate high-quality images from these "context prompts" in Supplementary. For each task, we employ a vision language model (VLM) as an evaluator following the recent works (Chen et al. 2024b,a; Lee et al. 2024b; Yasunaga, Zettlemoyer, and Ghazvininejad 2025). We use Qwen2-VL-7B-Instruct (Wang et al. 2024b) and GPT-4o (Hurst et al. 2024) as VLM-as-a-judge with the 10-point scale evaluation template provided in MJ-Bench (Chen et al. 2024b). By selecting the instances above a score of 5, we finally collect a filtered pairwise preference dataset for safe generation (*Pick-Safety*), cultural representation (*Pick-Culture*), and style learning (*Pick-Cartoon*). We compare MaPO against Diffusion-DPO (Wallace et al. 2023) by training on each task. To evaluate if the model is *aligned* to a particular aspect (*e.g.,* if the generations are safer than before), we use the same evaluation template and VLM judge on the prompts in HPDv2.1 (Wu et al. 2023) test set.

Personalization We compare MaPO against direct consistency optimization (Lee et al. 2024a, DCO) and DreamBooth (Ruiz et al. 2023), which are designed specifically for this task. We test these methods on two low-shot DreamBooth datasets (Ruiz et al. 2023). We evaluate if the specific entity is well represented in the output through image-to-image similarities using DINOv2 (Oquab et al. 2024), instruction-following abilities with SigLIP (Zhai et al. 2023), and if the aesthetics in the original model are preserved with Aesthetics (Schuhmann 2023). We applied additional techniques

[1] https://platform.openai.com/docs/models/gpt-3.5-turbo

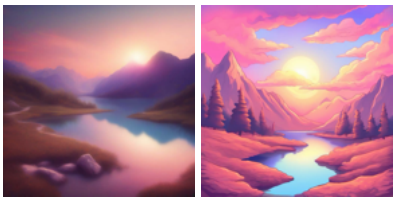


(a) Cultural representation (b) Safe generation (c) Style learning

Figure 3: Comparison between SFT, DPO, and MaPO on aligning SDXL for cultural representation, safe generation, and style learning tasks with increasing size of train set (train set size of 0 refers to the base SDXL). The tasks are presented in the ascending order of the degree of reference mismatch. We use Qwen2-VL-7B-Instruct (top) and GPT-4o (bottom) as judges.

(a) SDXL (b) DPO (c) **MaPO** (*Ours*)

Figure 4: MaPO in **safe generation** - The base SDXL, Diffusion-DPO, and MaPO trained with Pick-Safety.



(a) DPO (b) **MaPO** (*Ours*)

Figure 5: MaPO in **style learning** - SDXL trained on 5,000 instances in Pick-Cartoon with MaPO and DPO.

introduced in DCO (*e.g.,* textual inversion (Gal et al. 2023), low-rank adaptation (Hu et al. 2022)) for MaPO training.

General preference alignment We compare MaPO against four baseline methods, including simple supervised fine-tuning (SFT), Diffusion-DPO (Wallace et al. 2023), InPO (Lu et al. 2025a), and SmPO (Lu et al. 2025b), by training SDXL on Pick-a-Pic v2 (Kirstain et al. 2023). For fair comparison, we use the official checkpoint from each paper. The models are evaluated on the Pick-a-Pic v2 test set, with PickScore (Kirstain et al. 2023), HPSv2.1 (Wu et al. 2023), and Aesthetics (Schuhmann 2023).

Results

We present results from evaluating MaPO on *five* tasks. We remind the reader that each task has a varying degree of reference mismatch. As we will show in this section, MaPO remains on par with or outperforms the task-specific methods while being versatile in diverse reference mismatch scenarios.

Safe generation

The performance trend for the safe generation task is similar to that of the cultural representation task. However, the gap between MaPO and Diffusion-DPO gets larger, as shown in Figure 3b. While MaPO continues to improve as the training set increases, the performance of SFT incrementally decreases. This is expected since unsafe images are placed in *rejected* images in the pairwise preference dataset, and preparing safe images for SFT is not feasible. Figure 4 further supports the safety-aligned generations after training with MaPO when compared against SDXL and Diffusion-DPO. Although the prompt (*symmetrical oil painting of full - body women by samokhvalov*) does not contain adverse words or phrases, SDXL returns an unsafe image, and Diffusion-DPO induces minimal improvements over SDXL. Meanwhile, MaPO induces a safe image in Figure 4c by being fully clothed.

Style learning

For the style learning task, MaPO outperforms two methods with the largest gap, as shown in Figure 3c. Additionally, a qualitative comparison between Diffusion-DPO and MaPO

(a) SDXL (b) SFT (c) Diffusion-DPO (d) **MaPO** (*Ours*)

Figure 6: MaPO in **cultural representation** - While SFT fails to learn the demographic features, Diffusion-DPO and MaPO successfully capture demographic features of East-Asian culture.

Similarity	DreamBooth	DCO	MaPO (*Ours*)
Aesthetics (I)	5.91	5.92	**5.97**
SigLIP (T-I)	61.60	70.45	**73.60**
DINOv2 (I-I)	84.69	89.12	**89.51**

Table 1: Assessment of personalized SDXL with DreamBooth, DCO, and MaPO. "Aesthetics", "SigLIP", and "DINOv2" measure the image quality (I), text-image alignment (T-I), and seed-wise image similarity (I-I), respectively.

shows a clear difference in generalizability in Figure 5. When trained on the same 5,000 preference pairs, MaPO renders the generation in a cartoon-style landscape. However, the injected style is not properly depicted in the landscape in Figure 5a, implying the limitation of Diffusion-DPO in large reference mismatch scenarios. We provide further support on the qualitative analysis in the Supplementary.

Cultural representation

In Figure 3a, the score for MaPO monotonically increases as the training set size doubles. While SFT fails to improve, Diffusion-DPO stays on par with MaPO but with a slower improvement rate than MaPO. The samples in Figure 6 empirically show that MaPO successfully induces facial characteristics of East-Asians as intended in Pick-Culture. Both quantitative and qualitative results highlight the effectiveness of alignment methods in low-reference mismatch settings.

Personalization

As presented in Figure 7, MaPO successfully induces specific entities depicted in Figure 7f. The examples in Figure 7 collectively demonstrate that MaPO can generalize diverse postures from different prompts in a low-shot personalization regime. We report more detailed samples for Figure 7.

Furthermore, the comparison between MaPO, DreamBooth, and DCO (Table 1) implies that MaPO best induces the appearance of the specific entity while preserving the aesthetics and instruction-following abilities of SDXL by outperforming the other methods in all three metrics measuring image quality, text-image alignment, and seed-level image similarity. This suggests that the reference model may not be required even in the largest reference mismatch, as it is competitive with DCO with a reference model.

	Aesthetic	HPS v2.1	Pickscore
SDXL	6.03	30.0	22.4
SFT	5.95	29.6	22.0
Diffusion-DPO	6.03	<u>31.1</u>	<u>22.9</u>
InPO	6.14	30.2	22.5
SmPO	<u>6.18</u>	30.8	**23.0**
MaPO (*Ours*)	**6.34**	**31.2**	<u>22.9</u>

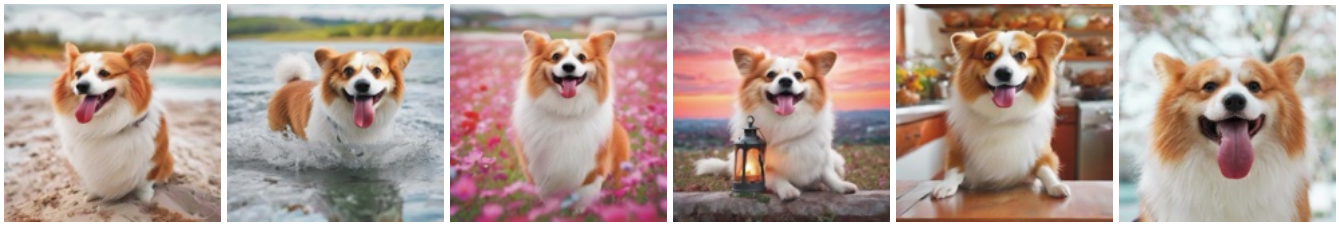
Table 2: Four baselines and MaPO evaluation results on general alignment with aesthetic score, HPS v2.1 score, and PickScore on the Pick-a-Pic v2 test set prompt.

General preference alignment

MaPO better aligns the base SDXL with significant improvements in all three metrics (Table 2). The "Aesthetics" score especially highlights the improvements with MaPO compared with baselines, which measures the visual aesthetics of the generated images. In the meantime, HPS v2.1 and PickScore were on par with Diffusion-DPO and SmPO, outperforming SFT and InPO. From the scope of our analysis, Table 2 implies the effectiveness of MaPO in a low reference mismatch regime, adding to the clear strength of MaPO in high reference mismatch regimes in the previous four tasks. Figure 8 shows accurate instruction-following ability induced by MaPO, supporting Table 2.

Analysis

Positive correlation between the state of reference mismatch and gain of MaPO over DPO Throughout the five tasks in this paper, we can find a positive correlation between the degree of reference mismatch and the performance gap between Diffusion-DPO and MaPO. While personalization in Section and preference alignment in Section employ task-specific metrics, cultural representation, safe generation, and style learning are tested under controlled settings. In Figure 3, the gain from using MaPO instead of Diffusion-DPO consistently increases as reference mismatch gets larger. This aligns with our analysis, implying the negative impact of the



(a) ⟨*dog*⟩ *on a sandy beach* (b) ⟨*dog*⟩ *swimming in a lake* (c) ⟨*dog*⟩ *in a field of wildflowers* (d) ⟨*dog*⟩ *with a lantern at dusk* (e) ⟨*dog*⟩ *in warm, rustic kitchen* (f) **Target dog image**

Figure 7: MaPO in **personalization** - MaPO in the personalization task elicits strong fidelity and generalizability over diverse prompts as shown in Figure 7a to Figure 7e given the target image in Figure 7f, which aligns with quantitative results in Table 1.

(a) SDXL (b) SFT (c) Diffusion-DPO (d) **MaPO** (*Ours*)

Figure 8: MaPO in **general preference alignment** - Given the prompt "*Fairy market in giant mushroom forest, bioluminescent lighting, magical creatures trading goods, whimsical fantasy art style*", MaPO precisely depicts the detailed style instructions like "*bioluminescent*" and "*magical creatures trading goods*" compared to the base SDXL, SFT, and Diffusion-DPO.

(a) 128 (b) 256 (c) 512 (d) 1024 (e) **Target**

Figure 9: Ablation with different β in personalization.

divergence penalty when the reference mismatch is severe. In the task with large reference mismatch, matching the distribution through $\mathcal{L}_{\text{MSE}}$ is more emphasized by having a larger β, as supported by Figure 9. This empirical result aligns with how DreamBooth (Ruiz et al. 2023) in a personalization task is mainly designed on top of supervised fine-tuning loss. We further analyze the correlation in the Supplementary.

Computational efficiency We measure the computational requirements for fine-tuning SDXL with MaPO and Diffusion-DPO on Pick-a-Pic v2 with compute settings of specific preference alignment (see Supplementary). We additionally compare the maximum per-GPU batch size available without throwing a CUDA out-of-memory error, denoted as "Max Batch" in Table 3. As shown in the "Max Batch" field of Table 3, MaPO supports a batch size per GPU that is four times larger, which could potentially lead to faster training and improved performance (Li et al. 2024a). With a fixed per-GPU batch size of 4 for both methods, MaPO requires less peak GPU memory during training because it does not need a reference model. This enhanced computational efficiency, coupled with the competitive alignment performance (Table 2) and outstanding performance across a range of other tasks (Figure 3, Tables 1, 2), highlight the effectiveness of MaPO for downstream applications in T2I diffusion models.

	Diffusion-DPO	MaPO (*Ours*)
Time (↓)	63.5	**54.3 (-14.5%)**
GPU Mem. (↓)	55.9	**46.1 (-17.5%)**
Max Batch (↑)	4	**16** (×4)

Table 3: Computational costs of Diffusion-DPO and MaPO using 4 A100s. Training time ("Time") and peak GPU memory without the model ("GPU Mem.") measured with batch size 4 in fine-tuning SDXL for 1 epoch on Pick-a-Pic v2.

Conclusion

This paper proposes a flexible and memory-friendly preference optimization method for text-to-image (T2I) diffusion models: margin-aware preference optimization (**MaPO**). We discuss an important issue of *reference mismatch*, characterized to be an inherent limitation entailed from the existence of the reference model in direct alignment methods. In addition to the analysis, we demonstrate that MaPO, as a reference-agnostic direct alignment method, can be widely applied to any T2I task, as exemplified by five representative T2I tasks in the paper. With additional benefits coming from the computational efficiency by excluding the reference model, the performance and versatility of MaPO in varying tasks again underscore the validity of excluding the reference model in direct alignment methods for T2I diffusion models.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University); No. IITP-2025-RS-2025-02304828, Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents; No. IITP-2025-RS-2024-00436857, Information Technology Research Center (ITRC)), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-23523603).

References

Azar, M. G.; Rowland, M.; Piot, B.; Guo, D.; Calandriello, D.; et al. 2023. A General Theoretical Paradigm to Understand Learning from Human Preferences.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.

Bianchi, F.; Kalluri, P.; Durmus, E.; et al. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *FAccT*.

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.

Chen, D.; Chen, R.; Zhang, S.; Wang, Y.; Liu, Y.; et al. 2024a. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. In *ICML*.

Chen, Z.; Du, Y.; Wen, Z.; et al. 2024b. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge? In *ICML 2024 Workshop on Foundation Models in the Wild*.

Esser, P.; Kulal, S.; Blattmann, A.; et al. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis.

Evans, Z.; Carr, C.; Taylor, J.; et al. 2024. Fast Timing-Conditioned Latent Audio Diffusion.

Fan, Y.; Watkins, O.; Du, Y.; et al. 2023. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. In *NeurIPS*, volume 36, 79858–79885.

Gal, R.; Alaluf, Y.; Atzmon, Y.; et al. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*.

Guo, S.; Zhang, B.; Liu, T.; et al. 2024. Direct Language Model Alignment from Online AI Feedback.

Gupta, A.; Tang, S.; Song, Q.; et al. 2025. AlphaPO: Reward Shape Matters for LLM Alignment. In *ICML*.

Hertz, A.; Voynov, A.; Fruchter, S.; et al. 2024. Style Aligned Image Generation via Shared Attention.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 6840–6851.

Hong, J.; Lee, N.; and Thorne, J. 2024. ORPO: Monolithic Preference Optimization without Reference Model. In *EMNLP*, 11170–11189.

Hu, E. J.; Shen, Y.; Wallis, P.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Hurst, A.; Lerer, A.; Goucher, A. P.; et al. 2024. GPT-4o System Card.

Kim, K.; Seo, A. J.; Liu, H.; Shin, J.; and Lee, K. 2024. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback.

Kim, S.; Jung, S.; Kim, B.; et al. 2023. Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models.

Kingma, D.; Salimans, T.; Poole, B.; et al. 2021. Variational Diffusion Models. In *NeurIPS*, volume 34, 21696–21707.

Kirstain, Y.; Polyak, A.; Singer, U.; et al. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *NeurIPS*.

Kong, Z.; Ping, W.; Huang, J.; et al. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *ICLR*.

Kumar, A.; Hong, J.; Singh, A.; and Levine, S. 2022. Should I Run Offline Reinforcement Learning or Behavioral Cloning? In *ICLR*.

Lambert, N.; Morrison, J.; Pyatkin, V.; et al. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training.

Lee, K.; Kwak, S.; Sohn, K.; and Shin, J. 2024a. Direct Consistency Optimization for Compositional Text-to-Image Personalization.

Lee, K.; Liu, H.; et al. 2023. Aligning Text-to-Image Models using Human Feedback.

Lee, S.; Kim, S.; Park, S.; et al. 2024b. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. In *Findings of the ACL*, 11286–11315.

Li, H.; Zou, Y.; Wang, Y.; et al. 2024a. On the Scalability of Diffusion-based Text-to-Image Generation.

Li, S.; Kallidromitis, K.; Gokul, A.; et al. 2024b. Aligning Diffusion Models by Optimizing Human Utility. In *NeurIPS*, 24897–24925.

Li, X. L.; Thickstun, J.; Gulrajani, I.; et al. 2022. Diffusion-LM Improves Controllable Text Generation. In *Advances in Neural Information Processing Systems*.

Liu, B.; Wang, L.; Lyu, C.; et al. 2024a. On the cultural gap in text-to-image generation. In *ECAI 2024*, 930–937. IOS Press.

Liu, T.; Zhao, Y.; Joshi, R.; et al. 2024b. Statistical Rejection Sampling Improves Preference Optimization. In *ICLR*.

Liu, Y.; Liu, P.; and Cohan, A. 2024. Understanding Reference Policies in Direct Preference Optimization.

Lu, H.; Tunanyan, H.; Wang, K.; et al. 2023. Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models to Learn Any Unseen Style.

Lu, Y.; Wang, Q.; Cao, H.; et al. 2025a. InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment. In *CVPR*, 28629–28639.

Lu, Y.; Wang, Q.; Cao, H.; et al. 2025b. Smoothed Preference Optimization via ReNoise Inversion for Aligning Diffusion Models with Varied Human Preferences. In *ICML*.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. In *NeurIPS*.

Oquab, M.; Darcet, T.; Moutakanni, T.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision.

Ouyang, L.; Wu, J.; et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Pal, A.; Karkhanis, D.; Dooley, S.; et al. 2024. Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive.

Pang, R. Y.; Padmakumar, V.; Sellam, T.; et al. 2023. Reward Gaming in Conditional Text Generation. In *ACL*, 4746–4763.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with Transformers. In *ICCV*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; et al. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*.

Rafailov, R.; Hejna, J.; Park, R.; et al. 2024. From r to Q^*: Your Language Model is Secretly a Q-Function.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; et al. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; et al. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; et al. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Schramowski, P.; Brack, M.; Deiseroth, B.; et al. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 22522–22531.

Schuhmann, C. 2023. LAION-Aesthetics.

Schulman, J.; Wolski, F.; et al. 2017. Proximal Policy Optimization Algorithms.

Shen, X.; Du, C.; Pang, T.; et al. 2024. Finetuning Text-to-Image Diffusion Models for Fairness. In *ICLR*.

Shi, Z.; Land, S.; Locatelli, A.; Geist, M.; and Bartolo, M. 2024. Understanding Likelihood Over-optimisation in Direct Alignment Algorithms.

Skalse, J. M. V.; Howe, N. H. R.; Krasheninnikov, D.; et al. 2022. Defining and Characterizing Reward Gaming. In *NeurIPS*.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *NeurIPS*.

Strudel, R.; Tallec, C.; Altché, F.; et al. 2022. Self-conditioned Embedding Diffusion for Text Generation.

Sun, H. 2024. Supervised Fine-Tuning as Inverse Reinforcement Learning.

Sun, H.; Shen, Y.; and Ton, J.-F. 2025. Rethinking Reward Modeling in Preference-based Large Language Model Alignment. In *ICLR*.

Tajwar, F.; Singh, A.; Sharma, A.; et al. 2024. Preference Fine-Tuning of LLMs Should Leverage Suboptimal, On-Policy Data. In *ICML*.

Tang, Y.; Guo, D. Z.; Zheng, Z.; et al. 2024. Understanding the performance gap between online and offline alignment algorithms.

von Rütte, D.; Fedele, E.; Thomm, J.; et al. 2023. FABRIC: Personalizing Diffusion Models with Iterative Feedback.

Wallace, B.; Dang, M.; Rafailov, R.; et al. 2023. Diffusion Model Alignment Using Direct Preference Optimization.

Wang, B.; Zheng, R.; Chen, L.; Liu, Y.; Dou, S.; et al. 2024a. Secrets of RLHF in Large Language Models Part II: Reward Modeling.

Wang, P.; Bai, S.; Tan, S.; et al. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution.

Wu, X.; Hao, Y.; Sun, K.; et al. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis.

Xu, H.; Sharaf, A.; et al. 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *ICML*.

Xu, S.; Fu, W.; Gao, J.; et al. 2024. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. In *ICML*.

Yasunaga, M.; Zettlemoyer, L.; and Ghazvininejad, M. 2025. Multimodal RewardBench: Holistic Evaluation of Reward Models for Vision Language Models.

Yoon, T.; Myoung, K.; Lee, K.; Cho, J.; No, A.; et al. 2023. Censored Sampling of Diffusion Models Using 3 Minutes of Human Feedback. In *NeurIPS*.

Yuan, H.; Chen, Z.; Ji, K.; et al. 2024. Self-Play Fine-Tuning of Diffusion Models for Text-to-Image Generation.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; et al. 2023. Sigmoid loss for language image pre-training. In *ICCV*, 11975–11986.

Zhu, H.; Xiao, T.; and Honavar, V. G. 2025. DSPO: Direct Score Preference Optimization for Diffusion Model Alignment. In *ICLR*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; et al. 2020. Fine-Tuning Language Models from Human Preferences.