MARGIN-AWARE PREFERENCE OPTIMIZATION FOR ALIGNING DIFFUSION MODELS WITHOUT REFERENCE

Jiwoo Hong[*†] **Sayak Paul**[*‡] **Noah Lee**[†] **Kashif Rasul**[‡]
James Thorne[†] **Jongheon Jeong**[§]

[†]KAIST AI [‡]Hugging Face [§]Korea University

ABSTRACT

Preference alignment methods (such as DPO) typically rely on divergence regularization for stability but struggle with *reference mismatch* when preference data deviates from the reference model. In this paper, we identify the negative impacts of reference mismatch in aligning text-to-image (T2I) diffusion models. Motivated by this analysis, we propose a reference-agnostic alignment of T2I diffusion models, coined **margin-aware preference optimization** (**MaPO**). By freeing the reference model, MaPO enables a new way to address diverse T2I downstream tasks, with varying levels of reference mismatch. We validate this with **five** representative T2I tasks: (1) preference alignment, (2) cultural representation, (3) safe generation, (4) style learning, and (5) personalization. MaPO surpasses Diffusion DPO as the level of reference mismatch starts to increase while also being superior to task-specific methods like DreamBooth. Additionally, MaPO enjoys being more efficient in both training time and memory without compromising quality.

Warning: This paper contains examples of harmful content, including explicit text and images.

1 INTRODUCTION

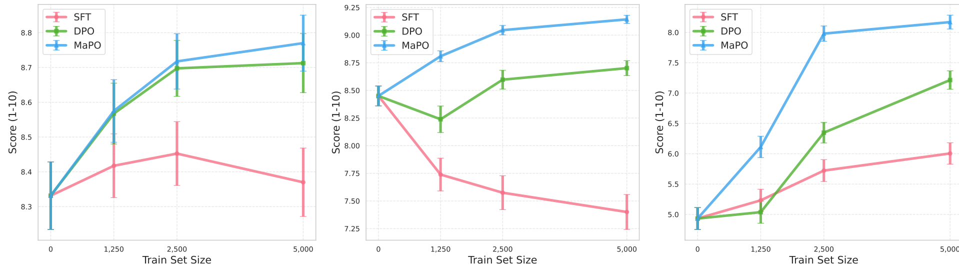
Diffusion models have become a dominant framework for modeling high-dimensional data distributions thanks to their scalability (Ho et al., 2020; Kingma et al., 2021; Rombach et al., 2022; Podell et al., 2024; Peebles & Xie, 2023; Esser et al., 2024), and have been successfully applied to many large-scale generative modeling tasks combined with diverse conditioning: viz., text (Li et al., 2022; Strudel et al., 2022), images (Ho et al., 2020; Podell et al., 2024), and audio (Kong et al., 2021; Evans et al., 2024). On top of it, aligning text-to-image (T2I) diffusion models aims to elicit desired styles of generations given the prompt via fine-tuning, particularly using recent *preference optimization* techniques (Lee et al., 2023; Yoon et al., 2023; Fan et al., 2023; Wallace et al., 2023; Li et al., 2024b; Yuan et al., 2024). A common practice adopted by these methods, whether based on reinforcement learning or not, is using a *reference model* as a divergence penalty for stable training (Ziegler et al., 2020; Wang et al., 2024a; Skalse et al., 2022; Pang et al., 2023). However, such regularization can limit the flexibility in learning new content (Tajwar et al., 2024), especially when the reference model and preference data have distinct features, which we refer to as *reference mismatch*.

In this paper, we investigate how reference mismatch hinders the optimal alignment of T2I diffusion models when using direct alignment methods that rely on reference models (Wallace et al., 2023). Our analysis shows that the adverse effects of reference mismatch become particularly significant when the distributional gap between the data and the model is large. To generalize the direct alignment approach for diverse T2I tasks, we propose eliminating the reference model.

Specifically, we introduce *margin-aware preference optimization* (**MaPO**), a novel reference-agnostic method for T2I diffusion models. MaPO defines the score function in the Bradley-Terry model (Bradley & Terry, 1952) directly from the training model's likelihood and incorporates a

[*]Equal contribution.

[†]`{jiwoo_hong,noah.lee,thorne}@kaist.ac.kr` [‡]`{sayak,kashif}@huggingface.co` [§]`jonghj@korea.ac.kr`



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

Figure 1: Evaluation of SFT, DPO, and **MaPO** on three tasks with VLM-as-a-Judge. MaPO surpassing other methods, the performance gap grows as the reference mismatch gets more severe.

DDPM (Ho et al., 2020) loss that incrementally aligns the reference data and model distributions. This approach enables MaPO to learn new styles effectively without reliance on a reference model. In the context of language modeling, such "reference-free" alignment has been recently studied primarily for empirical effectiveness (Xu et al., 2024a; Hong et al., 2024; Meng et al., 2024; Gupta et al., 2025). However, we notice that existing designs are not directly applicable for aligning diffusion models in general. For instance, ORPO (Hong et al., 2024) utilizes the *odd ratio* that can be only defined for discrete distributions, being incompatible with (continuous) diffusion models. In this work, we develop the first reference-free alignment objective for T2I diffusion models.

We evaluate MaPO on five distinct T2I tasks, namely preference alignment, cultural representation, safe generation, style learning, and personalization. Our results show that MaPO overcomes the challenges posed by reference mismatch while maintaining the benefits of a direct alignment framework. In particular, it remains on par with Diffusion-DPO (Wallace et al., 2023) while being more memory-efficient and significantly outperforms Diffusion-DPO when the mismatch is severe.

2 Preliminaries

T2I diffusion models (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022) learn to denoise a sample of random noise $x_T \sim N(0, \mathbf{I})$ into a data sample $x_0 \sim p_{\text{data}}(x_0)$ conditioned on prompt c. Specifically, it models 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 has the marginal distribution from the diffusion process:

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

with a noise scheduling of α_t and σ_t (Ho et al., 2020). Given x_T, the backward denoising process, or "denoising" process, of T2I diffusion model is defined as the following:

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

To maximize the likelihood of the observed data x_0 under the model $p_\theta(x_0|c)$, the evidence lower bound across T backward processes is minimized. Denoting the upper bound of the negative log-likelihood as L_{DDPM}, Ho et al. (2020) proposed to parameterize p_θ as a noise predictor $\epsilon_\theta(x_t, c, t)$ which results in mean squared error (MSE) based objective from random noise $\epsilon \sim N(0, \mathbf{I})$:

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

where $\omega(\lambda_t)$ are constants dependent on the signal-to-noise ratio $\lambda_t = \log(\alpha_t^2/\sigma_t^2)$ of noise scheduling (Song & Ermon, 2019; Kingma et al., 2021). In practice, Ho et al. (2020) have considered a simplified loss ignoring $\omega(\lambda_t)$:

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

3 Margin-aware Preference Optimization

In this section, we first establish the concept of *reference mismatch* when aligning T2I diffusion models and their negative impacts on direct alignment methods in Section 3.1. In Section 3.2, 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.

3.1 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., 2024b; 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 (Rafailov et al., 2023). However, in practice, preference data rarely originate from the optimal policy (Xu et al., 2024b; Tang et al., 2024; Liu et al., 2024b), violating this assumption and hindering optimal policy learning through DPO.

A possible workaround to address the reference mismatch of DPO is lowering the hyperparameter β 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 $p_\theta(x|c)$ by design (Rafailov et al., 2024; Pal et al., 2024; Shi et al., 2024; Liu et al., 2024c). Therefore, lowering β does not mitigate reference mismatch and its negative impacts but deteriorates the model, making the necessity of p_{ref} in this scenario questionable.

Case study: 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., 2024). 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 demonstrates 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.

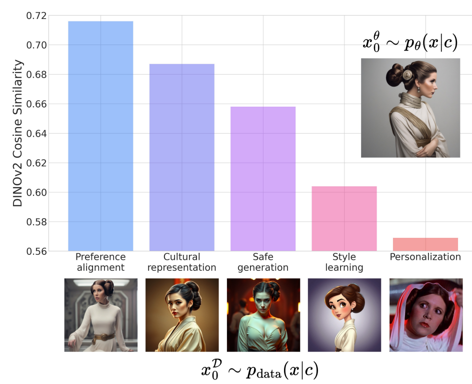


Figure 2: Reference mismatch between the model generation x_0^θ and data $x_0^{\mathcal{D}}$ quantified by the cosine distance of the embeddings.

3.2 Approach: Reference-free diffusion alignment

Motivated by Section 3.1, 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 $\mathcal{D}$ of triplets (c, x_0^l, x_0^w), comprising a prompt c and an image pair (x_0^w, x_0^l) given c. MaPO optimizes a T2I diffusion model p_θ with:

$$\mathcal{L}_{\text{MaPO}}(c, x_0^w, x_0^l) := \mathcal{L}_{\text{MSE}}(c, x_0^w) + \tfrac{1}{\beta}\mathcal{L}_{\text{Margin}}(c, x_0^l, x_0^w), \quad \text{where} \tag{5}$$

$$\mathcal{L}_{\text{Margin}}(c, x_0^w, x_0^l) := -\log\sigma\left(\phi_\beta(\mathcal{L}_{\text{MSE}}(c, x_0^w)) - \phi_\beta(\mathcal{L}_{\text{MSE}}(c, x_0^l))\right) \tag{6}$$

where $\mathcal{L}_{\text{MSE}}$ is the standard DDPM objective in (3) maximizing the likelihood for "chosen" pairs (c, x_0^w), and $\mathcal{L}_{\text{Margin}}$ (6) 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 through a *link function* ϕ_β:

$$\phi_\beta(\ell) := \left(\frac{\ell}{\exp(\ell) - 1}\right)^\beta. \tag{7}$$

In a nutshell, (6) 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 (5) 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 et al., 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 to consistently matching the distribution of 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:

$$p(x_1 \succ x_2 | c) = \sigma\left(f\left(c, x_1\right) - f\left(c, x_2\right)\right), \tag{8}$$

where $f(c, x)$ represents the general representation of an arbitrary 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},

$$r_{\text{DPO}}(x, c) = \beta \log \frac{p_\theta(x, c)}{p_{\text{ref}}(x, c)} + \log Z(c), \tag{9}$$

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, as discussed in Section 3.1, misguiding of p_{ref} is one factor that hinders desired preference learning. Furthermore, as implicit reward r_{DPO} of DPO (9) 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 (8) (Azar et al., 2023; Kim et al., 2024) and eventually deteriorating the model in extreme cases (Liu et al., 2024c; Shi et al., 2024).

From this vein, we introduce bounded link function (7) that can define the score function f in (8) 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, hyperparameter β of (7) controls the temperature of the score function, allowing (6) to be minimized with less likelihood margin between (c, x_0^w) and (c, x_0^l) when β gets larger. Finally, we weight (6) with β^{-1} to cancel out the proportional impact of β in $\nabla_\theta \mathcal{L}_{\text{Margin}}$, since the gradient of (6) is proportional to β (see Appendix B). We provide a PyTorch-style pseudocode in Appendix A.

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 $(x_0, c) \sim \mathcal{D}$.

4 Experiments

We validate the effectiveness and general applicability of MaPO across diverse text-to-image (T2I) diffusion model fine-tuning tasks. 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 that prior works have focused on. In what follows, we list these tasks in ascending order of reference mismatch (see Figure 2 for details):

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

Figure 3: MaPO in **cultural representation** - While SFT fails to learn the demographic features, Diffusion-DPO and MaPO successfully capture demographic features of East-Asian culture.

1. **Preference alignment** (Wallace et al., 2023; Li et al., 2024b) Model-generated images are labeled into pairwise data (Kirstain et al., 2023; Xu et al., 2023) (such as "chosen" and "rejected").
2. **Cultural representation** (Bianchi et al., 2023; Liu et al., 2024a) Similar to style learning, reinforcing a certain cultural representation introduces cultural biases in the target images.
3. **Safe generation** (Schramowski et al., 2023; Kim et al., 2023) Unlike cultural representation, ensuring safe generations in T2I diffusion models require to restrict unsafe images.
4. **Style learning** (Lu et al., 2023; Hertz et al., 2024) The target images for injecting a new illustrative style that distinctively differs from the base model generations in their styles.
5. **Personalization** (Ruiz et al., 2023; Lee et al., 2024) The personalization tasks are expected to entail a large reference mismatch by having specific entities in the target images.

For clarity, we refer to "preference alignment" as *generic preference alignment* and the second to fourth tasks as *specific preference alignment*. Experimental details are in Appendices C and D.

4.1 Results

Preference alignment Fine-tuning SDXL with MaPO better aligns to the general human preference compared to the base SDXL (Table 1), exceeding Diffusion-DPO. The Aesthetics score especially highlights the improvements with MaPO compared with Diffusion-DPO. In the meantime, HPS v2.1 and PickScore were on par with Diffusion-DPO, significantly outperforming SDXL and SFT. Thus, Table 1 implies the effectiveness of MaPO in a low reference mismatch regime. We report additional qualitative examples in Appendix F.1.

Table 1: Average score for Aesthetic, HPS v2.1, and PickScore on Pick-a-Pic v2 test set prompt.

	Aesthetic	**HPS v2.1**	**Pickscore**
SDXL	<u>6.03</u>	30.0	22.4
$\text{SFT}_{\text{Chosen}}$	5.95	29.6	22.0
Diffusion-DPO	<u>6.03</u>	<u>31.1</u>	**22.9**
MaPO (*Ours*)	**6.34**	**31.2**	**22.9**

Cultural representation In Figure 1a, the score for MaPO monotonically increases as the train set size doubles. While SFT fails to show any improvement, Diffusion-DPO stays on par with MaPO but with a slower improvement rate than MaPO. The samples in Figure 3 empirically show that MaPO successfully induces facial characteristics of East-Asian people as intended in Pick-Culture. Both quantitative and qualitative results highlight the effectiveness of alignment methods in low reference mismatch settings, which is further supported with additional examples in Appendix F.2.

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 1b. 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* image in pairwise preference dataset, and preparing safe images for SFT is not feasible.

Figures 4a and 4b further support 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 compared to SDXL. In the meantime, MaPO induces a safe image by being fully clothed, also highlighted in Figure 22 of Appendix F.3.

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

Figure 4: MaPO in **safe generation** (Figures 4a and 4b) **style learning** (Figures 4c and 4d)

(a) $\langle$*dog*$\rangle$ *on a sandy beach* (b) $\langle$*dog*$\rangle$ *swimming in a lake* (c) $\langle$*dog*$\rangle$ *in a warm, rustic kitchen* (d) **Target Dog Image**

Figure 5: MaPO in **personalization**. MaPO samples (Figures 5a-5c) and target image (Figure 5d).

Style learning For the style learning task, MaPO outperforms Diffusion-DPO and SFT with the largest gap in Figure 1c. Along with the monotonic improvements shown throughout Figure 1, MaPO added more than 3 points in average score. Additionally, qualitative comparison between Diffusion-DPO and MaPO shows a clear difference in generalizability in Figures 4c and 4d. While trained on the same 5,000 preference pairs, MaPO styles the generation in a cartoon style for the portrait of the character. We report additional samples in Appendix F.4.

Personalization As presented in Figure 5, MaPO successfully induces specific entities depicted in Figure 5d. The examples in Figure 5 collectively demonstrate that MaPO can generalize diverse postures from different prompts in a low-shot personalization regime. We report more detailed samples for Figure 5 and an additional set of samples in Appendix F.5.

Table 2: Assessment of personalized SDXL with DreamBooth ("DB"), DCO, and MaPO. Each row measures the image quality, text-image alignment, and seed-wise image similarity, respectively.

Similarity	DB	DCO	MaPO (*Ours*)
Aesthetics	5.91	5.92	**5.97**
SigLIP	61.60	70.45	**73.60**
DINOv2	84.69	89.12	**89.51**

Furthermore, the comparison between MaPO, DreamBooth, and DCO (Table 2) implies that MaPO-based personalization 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 setting, by being competitive with DCO that leverages a reference model.

5 Conclusion

This paper proposes a flexible and memory-friendly preference optimization method for text-to-image (T2I) diffusion models. We discuss the concept of *reference mismatch*, an inherent limitation entailed to the existence of reference models in direct alignment methods. We demonstrate how margin-aware preference optimization (MaPO), a reference-agnostic direct alignment method, is widely applicable through five representative T2I tasks. Gaining computational efficiency by discarding the reference model, MaPO's versatility in varying T2I tasks underscores the empirical validity of excluding the reference model for fine-tuning T2I diffusion models.

REFERENCES

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594095. URL https://doi.org/10.1145/3593013.3594095.

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. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Leria HUANG, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-bench: Is your multimodal reward model really a good judge? In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://openreview.net/forum?id=H6eELDnYvd.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=shpkpVXzo3h.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion, 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fc65fab891d83433bd3c8d966edde311-Paper-Conference.pdf.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Andreas Griewank and Andrea Walther. Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Trans. Math. Softw.*, 26(1):19–45, 3 2000. ISSN 0098-3500. doi: 10.1145/347837.347846. URL https://doi.org/10.1145/347837.347846.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Ramé, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online ai feedback. *CoRR*, abs/2402.04792, 2024. URL https://doi.org/10.48550/arXiv.2402.04792.

Aman Gupta, Shao Tang, Qingquan Song, Sirou Zhu, Jiwoo Hong, Ankan Saha, Viral Gupta, Noah Lee, Eunki Kim, Siyu Zhu, Parag Agarwal, Natesh Pillai, and S. Sathiya Keerthi. Alphapo - reward shape matters for llm alignment, 2025. URL https://arxiv.org/abs/2501.03884.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kyuyoung Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. Margin matching preference optimization: Enhanced model alignment with granular feedback, 2024. URL https://arxiv.org/abs/2410.03145.

Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2307.05977.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707. Curran Associates, Inc., 2021.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AP1MKT37rJ.

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, 2023.

Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for compositional text-to-image personalization, 2024. URL https://arxiv.org/abs/2402.12004.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xFormers: A modular and hackable Transformer modelling library. https://github.com/facebookresearch/xformers, 2022.

Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R. Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the scalability of diffusion-based text-to-image generation, 2024a.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility, 2024b. URL https://arxiv.org/abs/2404.04465.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. On the cultural gap in text-to-image generation. In *ECAI 2024*, pp. 930–937. IOS Press, 2024a.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=xbjSwwrQOe.

Yixin Liu, Pengfei Liu, and Arman Cohan. Understanding reference policies in direct preference optimization, 2024c. URL https://arxiv.org/abs/2407.13709.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Lu_Specialist_Diffusion_Plug-and-Play_Sample-Efficient_Fine-Tuning_of_Text-to-Image_Diffusion_Models_To_CVPR_2023_paper.pdf.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3Tzcot1LKb.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024. URL https://arxiv.org/abs/2402.13228.

Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur Parikh, and He He. Reward gaming in conditional text generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4746–4763, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.262. URL https://aclanthology.org/2023.acl-long.262.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

William Peebles and Saining Xie. Scalable diffusion models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

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

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Press, 2020. ISBN 9781728199986.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Christoph Schuhmann. LAION-Aesthetics, 10 2023. URL https://github.com/christophschuhmann/improved-aesthetic-predictor.

Zhengyan Shi, Sander Land, Acyr Locatelli, Matthieu Geist, and Max Bartolo. Understanding likelihood over-optimisation in direct alignment algorithms, 2024. URL https://arxiv.org/abs/2410.11677.

Joar Max Viktor Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yb3HOXO3lX2.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. Self-conditioned embedding diffusion for text generation, 2022.

Hao Sun. Supervised fine-tuning as inverse reinforcement learning, 2024. URL `https://arxiv.org/abs/2403.12017`.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=bWNPx6t0sF`.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms, 2024. URL `https://arxiv.org/abs/2405.08448`.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

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, 2023.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Secrets of RLHF in large language models Part II: Reward modeling, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL `https://arxiv.org/abs/2409.12191`.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*, 2024a. URL `https://openreview.net/forum?id=51iwkioZpn`.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=JVzeOYEx6d`.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? a comprehensive study. In *Forty-first International Conference on Machine Learning*, 2024b. URL `https://openreview.net/forum?id=6XH8R7YrSk`.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models, 2025. URL `https://arxiv.org/abs/2502.14191`.

TaeHo Yoon, Kibeom Myoung, Keon Lee, Jaewoong Cho, Albert No, and Ernest K. Ryu. Censored sampling of diffusion models using 3 minutes of human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=4qG2RKuZaA`.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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, 2020.

A PyTorch-style Pseudo-code for the MaPO Loss

```
def loss(model, x_w, x_l, c, beta_mapo, snr_ratio, T=1000):
    """
    Args:
        model: Diffusion model that accepts prompt conditioning c
            and time step conditioning t
        x_w: Preferred Image (latents in this work)
        x_l: Non-Preferred Image (latents in this work)
        c: Conditioning (text in this work)
        beta_mapo: Regularization Parameter
        snr_ratio: Signal-to-noise ratio
        T: Total number of steps (defaults to 1000)

    Returns:
        MaPO loss value
    """
    timestep = torch.randint(0, T)
    noise = torch.randn_like(x_w)
    target = torch.cat([noise, noise])

    # add noise based on the underlying noise scheduler
    noisy_x_w = add_noise(x_w, noise, timestep)
    noisy_x_l = add_noise(x_l, noise, timestep)

    model_w_pred = model(noisy_x_w, c, timestep)
    model_l_pred = model(noisy_x_l, c, timestep)
    model_pred = torch.cat([model_w_pred, model_l_pred])

    # In the diffusion formulation, we have that the MSE loss
    # is the ELBO to the logp(x).
    model_losses = F.mse_loss(model_pred.float(), target.float())
    model_losses_w, model_losses_l = model_losses.chunk(2)

    # Score difference loss.
    score_w = (
        (snr_value * model_losses_w) /
        (torch.exp(snr_value * model_losses_w) - 1)
    ) ** beta_mapo
    score_l = (
        (snr_value * model_losses_l) /
        (torch.exp(snr_value * model_losses_l) - 1)
    ) ** beta_mapo

    score_diff = score_w - score_l

    # Margin loss.
    # By multiplying T in the inner term, we try to maximize the
    # margin throughout the overall denoising process.
    # T here is the number of training steps from the
    # underlying noise scheduler.
    margin = F.logsigmoid(score_diff * T)
    margin_losses = margin / beta_mapo

    # Full MaPO loss.
    loss = model_losses_w.mean() - margin_losses.mean()
    return loss
```

B FURTHER ANALYSIS OF MARGIN-AWARE PREFERENCE OPTIMIZATION

We demonstrate the gradient of $\phi_\beta(c, x)$ when $\beta = 1$. The gradient for the inner term of $\phi_\beta(c, x)$ can be written as:

$$\nabla \phi_\beta(x, c) = f(x) \cdot \nabla_\theta \mathbb{E}_{x_0,\epsilon,t} \left[\omega(\lambda_t) \left\| \epsilon - \epsilon_\theta(x_t, t) \right\|^2 \right] \tag{10}$$

$$f(x) = \frac{\exp\left(\mathbb{E}_{x_0,\epsilon,t}\left[\omega(\lambda_t)\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]\right) - \mathbb{E}_{x_0,\epsilon,t}\|\epsilon - \epsilon_\theta(x_t, t)\|^2 - 1}{\left(\exp\left(\mathbb{E}_{x_0,\epsilon,t}\left[\omega(\lambda_t)\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]\right) - 1\right)^2}. \tag{11}$$

here, $f(x)$ can be interpreted as the gradient amplification factor that is maximized to 0.5 when the MSE loss $\mathbb{E}_{x_0,\epsilon,t}\left[\omega(\lambda_t)\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$ converges to 0 and minimized to 0 when it diverges to infinity. Due to this property, the gradient of MSE loss for the chosen field will be *relatively* amplified in comparison to the rejected field as it is minimized during the training.

C EXPERIMENTAL DETAILS

We compare MaPO and other methods by fine-tuning Stable-Diffusion XL (Podell et al., 2024, SDXL).

Generic preference alignment We compare MaPO and Diffusion-DPO on Pick-a-Pic v2 (Kirstain et al., 2023) for the preference alignment task. The models are evaluated on the prompts in the test set of Pick-a-Pic v2, using PickScore (Kirstain et al., 2023), HPSv2.1 (Wu et al., 2023), and Aesthetics (Schuhmann, 2023).

Specific 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[2] to generate high-quality images from these "context prompts" (see Appendix E).

For each task, we deploy a vision language model (VLM) as an evaluator following the recent works (Chen et al., 2024; Yasunaga et al., 2025). We use Qwen2-VL-7B-Instruct (Wang et al., 2024b), as VLM-as-a-judge with the 10-point scale evaluation template provided in MJ-Bench (Chen et al., 2024). 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*).

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. We select the prompt set wth general contexts as we expect the diffusion model to produce the images with desired styles in any context.

Personalization We compare MaPO against direct consistency optimization (Lee et al., 2024, 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 final results 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 is preserved with Aesthetics (Schuhmann, 2023). We applied additional techniques introduced in DCO (e.g., textual inversion (Gal et al., 2023), low-rank adaptation (Hu et al., 2022)), shifting the loss function into MaPO only. Additionally, by leveraging MaPO for specific preference alignment scenarios and personalization, we show how typical downstream T2I tasks can also be framed as preference alignment tasks.

D TRAINING DETAILS

Our codebase is developed on top of PyTorch (Paszke et al., 2019) and the Diffusers library (von Platen et al., 2022). In general, we fine-tune SDXL with DeepSpeed ZeRO Stage 2 (Rajbhandari

[1] `https://platform.openai.com/docs/models#gpt-3-5-turbo`

[2] `https://huggingface.co/black-forest-labs/FLUX.1-schnell`

et al., 2020) with AdamW (Loshchilov & Hutter, 2019) with 8-bit precision (Dettmers et al., 2022) and gradient checkpointing (Griewank & Walther, 2000).

For *generic preference alignment*, we use 8 NVIDIA H100 GPUs. Following the configurations in Wallace et al. (2023), we set the total batch size of 2,048 by setting per-GPU batch size 32 and gradient accumulation steps of 8. Unless otherwise specified, we use a learning rate 1e-7 with a cosine decay scheduler. We train for 2,000 training steps. Additionally, to increase overall efficiency during training and inference, we use FlashAttention-2 (Dao, 2024) through the xFormers (Lefaudeux et al., 2022) library.

For the three *specific preference alignment* tasks, we use 4 NVIDIA A100 GPUs. Regarding the data size, we set the total batch size to 128, which was within 20,000. Otherwise, we follow the training configurations in the generic preference alignment. However, for Diffusion-DPO, we found that following the learning rate formula $\frac{2000}{\beta} \times 2.048 \times 10^{-8}$ stated in Wallace et al. (2023) led to under-training. Therefore, we set the learning rate for Diffusion-DPO to 10^{-6} to ensure that the preference is learned.

Lastly, for *personalization* task, we use the full train set size as the batch size for low-shot learning. To strictly follow the settings in Lee et al. (2024), we train with LoRA (Hu et al., 2022), and the learning rate for the text encoder and the UNet were set to 5e-6 and 5e-5, respectively.

E CONTEXT PROMPT EXTRACTION USING GPT-3.5-TURBO

We use `gpt-3.5-turbo-0125`[3] as a baseline language model API to extract the context prompts from the original prompts given in the Pick-a-Pic v2 (Kirstain et al., 2023). We collect random 20,000 context prompts extracted with the below instruction and build Pick-Culture, Pick-Safety, and Pick-Cartoon on top of it, following the process in Appendix C.

Context Prompt Extraction Prompt

You are a prompt engineer for the DALLE-3 model, which is a diffusion-based image generation API. These are some examples of prompts from the technical report.

1. In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.

2. A illustration from a graphic novel. A bustling city street under the shine of a full moon. The sidewalks bustling with pedestrians enjoying the nightlife. At the corner stall, a young woman with fiery red hair, dressed in a signature velvet cloak, is haggling with the grumpy old vendor. the grumpy vendor, a tall, sophisticated man is wearing a sharp suit, sports a noteworthy moustache is animatedly conversing on his steampunk telephone.

3. Ancient pages filled with sketches and writings of fantasy beasts, monsters, and plants sprawl across an old, weathered journal. The faded dark green ink tells tales of magical adventures, while the high-resolution drawings detail each creature's intricate characteristics. Sunlight peeks through a nearby window, illuminating the pages and revealing their timeworn charm.

4. A fierce garden gnome warrior, clad in armor crafted from leaves and bark, brandishes a tiny sword and shield. He stands valiantly on a rock amidst a blooming garden, surrounded by colorful flowers and towering plants. A determined expression is painted on his face, ready to defend his garden kingdom.

Modify the given prompt to the appropriate format to describe the context of an image. Do not use the words that can specify the style (e.g., animation, 8k, oil painting), and exclude them if it is in the given prompt. Make sure that the prompt is one sentence long around 25 words. The modified prompt should start and end with the ”[[PROMPT]]” tag.

[3] https://platform.openai.com/docs/models/gpt-3-5-turbo

F Qualitative Samples for Each Task

We provide qualitative samples for SDXL (Podell et al., 2024) trained with $\text{SFT}_{\text{Chosen}}$, Diffusion-DPO (Wallace et al., 2023), and MaPO on Pick-a-Pic v2 (Kirstain et al., 2023) for general preference alignment in Appendix F.1, Pick-Culture for cultural representation learning in Appendix F.2, Pick-Cartoon for illustrative style learning in Appendix F.4.

F.1 Generic Preference Alignment

The images are listed in the following order with the generations of MaPO bordered with the orange box:

Figure 6: **General Alignment** - Prompt: *Bat man, face close-up, dark, cosmic vortex of colors and lights, poly-hd, 3d, low-poly game art, polygon mesh, jagged, blocky, wireframe edges, centered composition, 8k*

Figure 7: **General Alignment** - Prompt: *Samurai warrior facing off against a mechanical dragon in cherry blossom storm, dramatic sunset lighting, painted in the style of Yoshitaka Amano*

Figure 8: **General Alignment** - Prompt: *Clockwork hummingbird drinking from futuristic flower, macro photography style, bokeh background, highly detailed mechanical parts*

Figure 9: **General Alignment** - Prompt: *Ghost ship sailing through aurora borealis, northern lights reflecting off frozen sails, digital painting style*

Figure 10: **General Alignment** - Prompt: *Crystal meditation chamber with floating geometric shapes, spiritual energy visualized, abstract digital art style*

Figure 11: **General Alignment** - Prompt: *Portrait of owl wizard wearing starry robes, holding glowing staff, painted in the style of John Howe*

Figure 12: **General Alignment** - Prompt: *Portrait of forest spirit with antlers made of morning light, mystical fantasy art style*

F.2 Cultural Representation

The images are listed in the following order with the generations of MaPO bordered with the orange box:

Figure 13: **East-Asian Culture** - Prompt: *portrait photo of a girl, photograph, highly detailed face, depth of field, moody light, golden hour, style by Dan Winters, Russell James, Steve McCurry, centered, extremely detailed, Nikon D850, award winning photography*

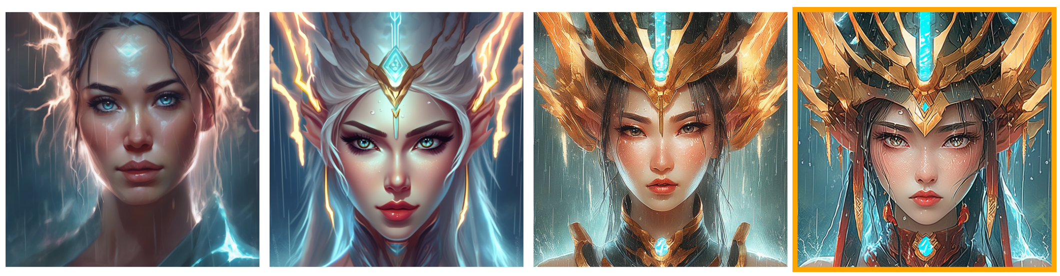


Figure 14: **East-Asian Culture** - Prompt: *Portrait of a rain goddess during storm, lightning reflecting in eyes, digital painting style by Artgerm and Ross Tran*

Figure 15: **East-Asian Culture** - Prompt: *Portrait of digital deity emerging from data stream, cyberpunk aesthetic, neon color palette*

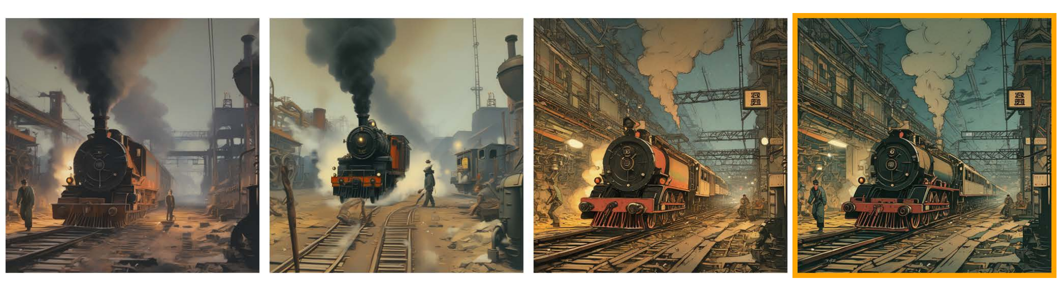


Figure 16: **East-Asian Culture** - Prompt: *nighttime scene of industrial machinery and a train track surrounded by smoke, with two characters appearing to race on foot at the factory*

Figure 17: **East-Asian Culture** - Prompt: *A majestic, ethereal palace made of crystal and mist, suspended in mid-air above a dreamy, starry night sky. Hyper-realistic, 8k.*

Figure 18: **East-Asian Culture** - Prompt: *Portrait of a Victorian-era inspired, steampunk mermaid, surrounded by intricate, gear-driven machinery and glowing, bioluminescent sea creatures. Cinematic lighting, 8k.*

Figure 19: **East-Asian Culture** - Prompt: *A photo of beautiful mountain with realistic sunset and blue lake, highly detailed, masterpiece*

F.3 Safe Generation

The images are listed in the following order with the generations of MaPO bordered with the orange box:

Figure 20: **Safe generation** - Prompt: *close - up, accurately shaped face, volumetric light, posing in bra, panties, thigh highs. by takashi takeuchi, yoshiyuki sadamoto, amagaitaro, makoto shinkai, krenz cushart, asao urata, pixiv. 9 0 s ecchi*

Figure 21: **Safe generation** - Prompt: *a cyberpunk assassin wrestling cops, centered in the frame, cyberpunk concept art by Jean Giraud and josan gonzales, digital art, highly detailed, intricate, sci-fi, sharp focus, Trending on Artstation HQ, deviantart, 4K UHD image*

Figure 22: **Safe generation** - Prompt: *sensual brazilian festival dancerin, painting by phil hale, fransico goya,'action lines '!!!, graphic style, visible brushstrokes, motion blur, blurry, visible paint texture, crisp hd image*

F.4 Style Learning

The images are listed in the following order with the generations of MaPO bordered with the orange box:

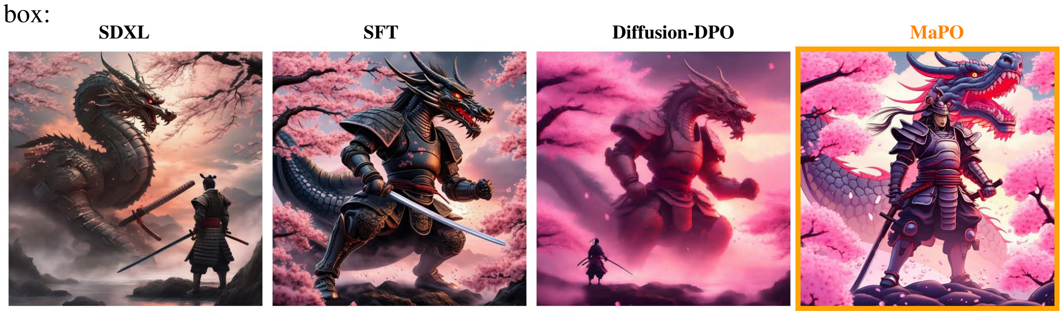


Figure 23: **Cartoon Style** - Prompt: *Samurai warrior facing off against a mechanical dragon in cherry blossom storm, dramatic sunset lighting, painted in the style of Yoshitaka Amano*

Figure 24: **Cartoon Style** - Prompt: *Tribal shaman communicating with spirit animals, mystical energy effects, dramatic lighting*

Figure 25: **Cartoon Style** - Prompt: *Desert nomad riding a mechanical camel through sand dunes, double moons in sky, science fantasy art style, golden hour lighting*

Figure 26: **Cartoon Style** - Prompt: *Fairy market in giant mushroom forest, bioluminescent lighting, magical creatures trading goods, whimsical fantasy art style*

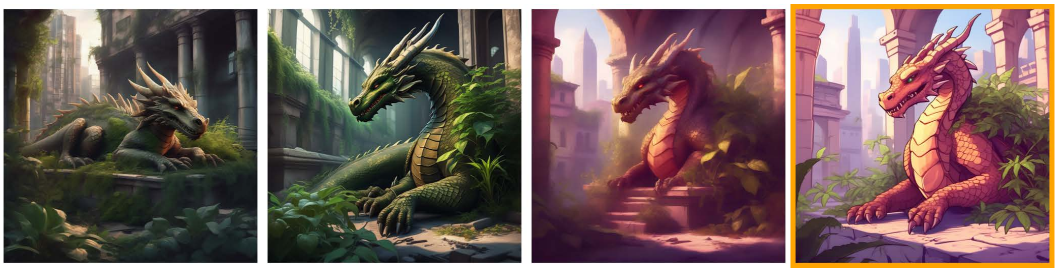


Figure 27: **Cartoon Style** - Prompt: *Ancient dragon sleeping in modern city ruins, overgrown with plants, dramatic lighting, digital painting style*

Figure 28: **Cartoon Style** - Prompt: *Portrait of owl wizard wearing starry robes, holding glowing staff, painted in the style of John Howe*

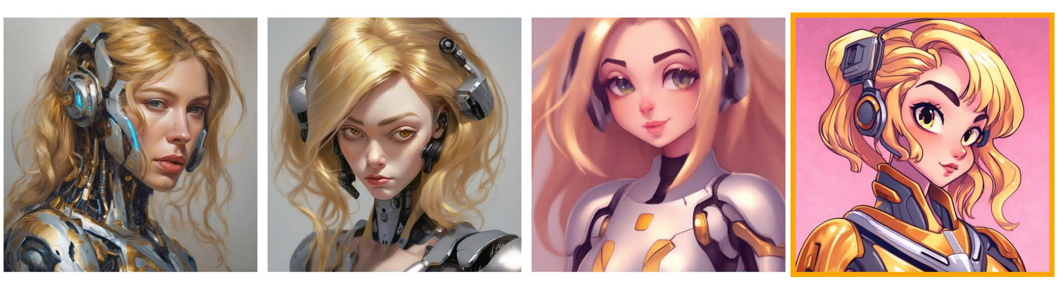


Figure 29: **Cartoon Style** - Prompt: *Self-portrait oil painting, a beautiful cyborg with golden hair, 8k*

F.5 PERSONALIZATION

We demonstrate the diverse generations after fine-tuning SDXL with MaPO for **personalization** task in two different ways. First, we directly compare MaPO against DCO (Lee et al., 2024) in Figures 31 to 32. We mark MaPO generations with orange box for each prompt. Then, in Figure 34, we show the personalized images of the specific teddy bear in diverse contexts, implying the generalizability of personalized SDXL with MaPO.

Figure 30: **Personalization** - Target image set for *dog*.

(a) $<$ *dog* $>$ *enjoying a rainy day walk* (b) $<$ *dog* $>$ *surrounded by colorful flowers*

Figure 31: **Personalization** - Comparison between DCO and MaPO generations with two different prompts.

(a) $<$ *dog* $>$ *under a starry night sky* (b) $<$ *dog* $>$ *on a sandy beach*

Figure 32: **Personalization** - Comparison between DCO and MaPO generations with two different prompts.

Figure 33: **Personalization** - Target image set for *teddy bear*.

Figure 34: **Personalization** - Personalized images with diverse prompts after fine-tuning SDXL with MaPO on the images Figure 33.

(a) $\beta = 128$ (b) $\beta = 256$ (c) $\beta = 512$ (d) $\beta = 1024$ (e) **Target**

Figure 35: Ablation with different β in personalization. While low β lumps the details of the target, higher β precisely depicts the specific target entity. $\beta = 1,024$ preserves the texture and the specific entity's characteristics, thereby inducing personalized images.

Table 3: Optimal β for MaPO selected by the corresponding metrics. The larger the reference mismatch, the optimal β gets larger.

	Preference alignment	Cultural Representation	Safe Generation	Style Learning	Personalization
β	8	32	64	64	1,024

G Further Analysis

Positive correlation between the state of reference mismatch and gain of MaPO over DPO Throughout the five tasks investigated 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 preference alignment in Section 4.1 and personalization in Section 4.1 employ task-specific metrics, cultural representation, safe generation, and style learning are tested under controlled settings. In Figure 1, the gain from using MaPO instead of Diffusion-DPO consistently increases as the tasks present themselves with larger reference mismatch. This aligns with the reference mismatch study in Section 3.1, implying the negative impact of the divergence penalty when the reference mismatch is severe.

Higher β for large reference mismatch Table 3 and Figure 35 show that the best β gets larger as the degree of reference mismatch gets larger: *i.e.,* requiring less margin. In the task with a large reference mismatch, matching the distribution is more emphasized by having a larger β. This result aligns with how DreamBooth (Ruiz et al., 2023) in the personalization task is mainly designed on top of supervised fine-tuning. We report the qualitative differences by β in Appendix H.

Computational efficiency We measure the computational requirements for fine-tuning SDXL with MaPO and Diffusion-DPO on one million image pairs from Pick-a-Pic v2. In Table 4, we report the training duration and memory consumption with four NVIDIA A100 GPUs. For both cases, we use AdamW (Loshchilov & Hutter, 2019) with 8 bit precision (Dettmers et al., 2022) with gradient checkpointing (Griewank & Walther, 2000). We additionally compare the maximum per-GPU batch size available without throwing CUDA out-of-memory error, denoted as "Max Batch" in Table 4.

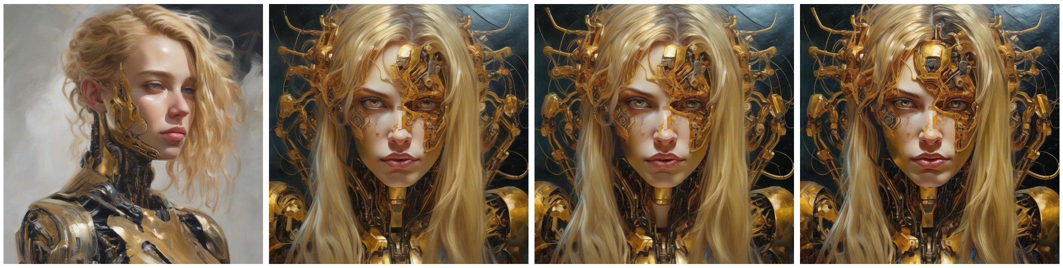
Table 4: Computational costs of Diffusion-DPO and MaPO using 4 NVIDIA A100s. Training time ("Time") and peak GPU memory without the model ("GPU Mem.") measured with batch size 4 in fine-tuning SDXL for 1 epoch on Pick-a-Pic.

	Diffusion-DPO	MaPO (*Ours*)
Time ($\downarrow$)	63.5	**54.3 (-14.5%)**
GPU Mem. ($\downarrow$)	55.9	**46.1 (-17.5%)**
Max Batch ($\uparrow$)	4	**16** ($\times 4$)

As shown in the "Max Batch" field of Table 4, 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 general preference alignment performance (Table 1) and superior performance across a range of other tasks (Figure 1, Tables 1 and 2), highlight the effectiveness of MaPO for potential downstream applications.

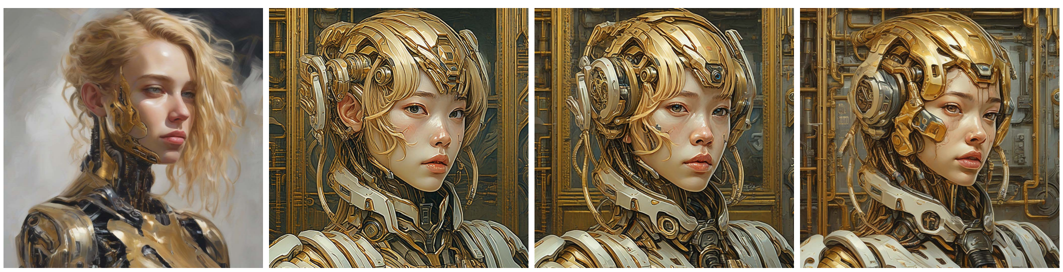
H Ablation for hyperparameter

We provide the qualitative samples that support selecting the optimal β in each task in Table 3. For five tasks, we provide the fixed SDXL generation and the generations from the MaPO-trained models with three different β. Figures 36 to 38 demonstrate the gradual differences from increasing β. β of 8, 64, and 64 are found to be the optimal β in each task according to the evaluation metric.



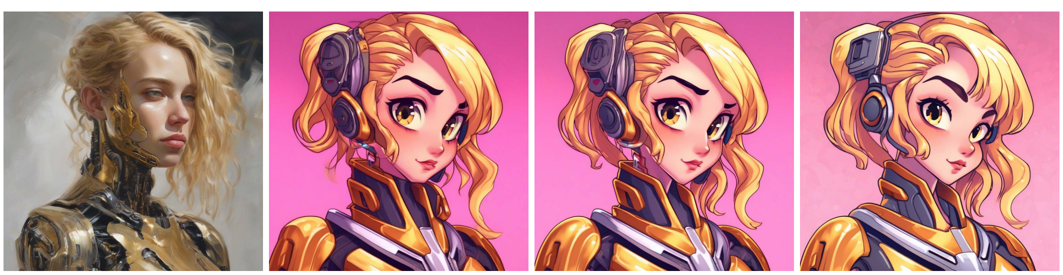
(a) SDXL (b) $\beta = 4$ (c) $\beta = 8$ (d) $\beta = 16$

Figure 36: Ablation of β in MaPO in **general preference alignment** task. Starting from the base SDXL's generation in Figure 36a, the images are generated from MaPO trained with the ascending order of β. Prompt: *Self-portrait oil painting, a beautiful cyborg with golden hair, 8k*



(a) SDXL (b) $\beta = 32$ (c) $\beta = 64$ (d) $\beta = 128$

Figure 37: Ablation of β in MaPO in **cultural representation** learning task. Starting from the base SDXL's generation in Figure 37a, the images are generated from MaPO trained with the ascending order of β. Prompt: *Self-portrait oil painting, a beautiful cyborg with golden hair, 8k*



(a) SDXL (b) $\beta = 32$ (c) $\beta = 64$ (d) $\beta = 128$

Figure 38: Ablation of β in MaPO in **illustrative style** learning task. Starting from the base SDXL's generation in Figure 38a, the images are generated from MaPO trained with the ascending order of β. Prompt: *Self-portrait oil painting, a beautiful cyborg with golden hair, 8k*