Free Lunch Alignment of Text-to-Image Diffusion Models without Preference Image Pairs

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in diffusion-based text-to-image (T2I) models have led to remarkable success in generating high-quality images from textual prompts. However, ensuring accurate alignment between the text and the generated image remains a significant challenge for state-of-the-art diffusion models. To address this, existing studies often employ reinforcement learning with human feedback (RLHF) to align T2I outputs with human preferences. These methods, however, either rely directly on paired image preference data or require a learned reward function, both of which depend heavily on costly, high-quality human annotations and thus face scalability limitations. In this work, we introduce **Text Preference Optimization (TPO)**, a novel framework that enables "free-lunch" alignment of T2I models, achieving alignment without the need for paired image preference data. TPO works by training the model to prefer matched prompts over mismatched prompts, which are constructed by perturbing original captions using a large language model (LLM). Our framework is general and compatible with existing preference-based algorithms. We extend both DPO and KTO to our setting, resulting in **TDPO** and **TKTO**. Quantitative and qualitative evaluations across multiple benchmarks show that our methods consistently outperform their original counterparts, yielding superior human preference scores and better text-to-image alignment.

1 Introduction

Text-to-Image (T2I) models, driven by diffusion Karras et al. (2022); Kingma et al. (2023); Rombach et al. (2022a); Song et al. (2020), rectified flow (Lipman et al., 2022; Liu et al., 2022), and next-token prediction methods (Wang et al., 2024), have seen significant advances in generating high-quality images from textual descriptions. This is achieved by pretraining on large-scale, high-quality image caption datasets. However, the trained model's performance relies heavily on the quality of the pretraining datasets, and does not necessarily reflect human preferences for image quality and textual alignment. To mitigate this, recent studies have focused on aligning text-to-image (T2I) models with human preferences in the post-training phase.

Motivated by the success of large language models (LLMs), recent work has introduced reinforcement learning from human feedback (RLHF) for text-to-image (T2I) alignment (Black et al., 2024; Clark et al., 2024; Fan et al., 2023; Lee et al., 2023). These methods typically train a reward model to predict human preferences and then optimize the generator with policy-gradient updates. Despite promising results, the pipeline is complex—reward modeling can be unstable—and resource-intensive due to the need for substantial preference-labeled data. Moreover, the high-quality datasets used for initial training offer limited benefit at this stage, since preference alignment still requires a separate and costly annotation effort.

More recently, Direct Preference Optimization (DPO)(Rafailov et al., 2023) and Kahneman–Tversky Optimization (KTO)(Ethayarajh et al., 2024) simplify preference learning via a single-stage, closed-form objective, and have been extended to T2I models (Li et al., 2024; Wallace et al., 2024). However, these approaches still rely heavily on high-quality preference-labeled image pairs, which are expensive to collect and can be sensitive to noise and inconsistencies in human annotations (Yang et al., 2025).

We therefore ask: *Can we improve text-to-image model alignment without annotating human preference image pairs?*

Motivated by contrastive methods used in training vision-language models, we propose to leverage the high-quality datasets originally used for T2I training by optimizing preference alignment over text pairs rather than image pairs, as in Diffusion-DPO or Diffusion-KTO. This approach is based on the observation that, *given an image-caption dataset, generating text preference pairs for a given image is significantly easier than constructing image preference pairs for a single prompt.* To enable the construction of such text-preference pairs, we utilize LLMs to generate negative samples—prompts that closely resemble the original but are intentionally mismatched. For example, as shown in Fig. 1, our pipeline flips the word "inside" in the original prompt to "outside", inducing a spatial layout change. This strategy encourages the model to focus on subtle distinctions in prompt semantics, resulting in improved alignment performance. Our contributions are threefold:

- **Preference-data-free alignment.** We propose a novel method for aligning text-to-image diffusion models without requiring human preference data, offering a "free lunch" post-training solution while improving text-to-image alignment.
- **Generalizable pipeline.** Our approach is model-agnostic and can be seamlessly integrated into any RLHF-based method that utilizes preference pairs, making it broadly applicable across existing alignment frameworks.
- **State-of-the-art results.** By adapting DPO and KTO into our framework—TDPO and TKTO—we achieve state-of-the-art performance both qualitatively and quantitatively, surpassing the baselines without using any human preference annotations.

2 Related Work

Text-to-Image Models. Text-to-image diffusion models have emerged as one of the most powerful and widely adopted generative techniques for image synthesis (Balaji et al., 2022; Chang et al., 2023; Ho et al., 2020; Karras et al., 2022; Kingma et al., 2023; Rombach et al., 2022a; Shi et al., 2020; Song et al., 2020). These models are capable of generating high-quality images that closely match the semantics of a given text prompt. Despite their impressive performance, achieving precise alignment between textual descriptions and visual outputs remains a key challenge. Recent works have explored improvements through various directions, such as enhanced text encoding (Liu et al., 2024b; Ma et al., 2024; Wu et al., 2023a), improved text embedding interaction architectures (Esser et al., 2024a; Liu et al., 2024a), better image captioning (Betker et al., 2023; Lei et al., 2023), and inference-time strategies (Jiang et al., 2024; Prabhudesai et al., 2024; Shen et al., 2024; Wallace et al., 2023). Orthogonal to these approaches, our work focuses on improving text-image alignment through a novel preference optimization strategy, providing a complementary direction to existing methods.

Preference Alignment Foundations in LLMs. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2023; Ouyang et al., 2022) established the core paradigm of aligning generative models with human preferences by learning from preference annotations. More recent methods, especially Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants such as KTO and SPPO (Ethayarajh et al., 2024; Shao et al., 2024; Wu et al., 2024), remove explicit reward-model training and optimize preferences more directly. In this work, we use these methods mainly as optimization foundations and adapt them to text-to-image alignment, where the objective is to improve human preference over generated images. We build on the optimization strategies pioneered in language alignment and adapt them to the multimodal setting.

Text-to-Image Preference Optimization. As with LLMs, text-to-image (T2I) diffusion models must also be aligned to respect user preferences and safety constraints. Building on the success of preference-based methods such as DPO, several studies have recently transplanted these objectives to the T2I setting and reported strong gains (Karthik et al., 2024; Li et al., 2024; Miao et al., 2024; Wallace et al., 2024; Yang et al., 2024). In particular, *Diffusion-DPO* (Wallace et al., 2024) and *Diffusion-KTO* (Li et al., 2024) extend the DPO and KTO objectives (Ethayarajh et al., 2024) to diffusion-based generators. Recent works have explored preference-based alignment for diffusion models. InPO Lu et al. (2025a) assigns rewards to specific latent variables using inversion. SmPO-Diffusion Lu et al. (2025b) models graded human preferences through a

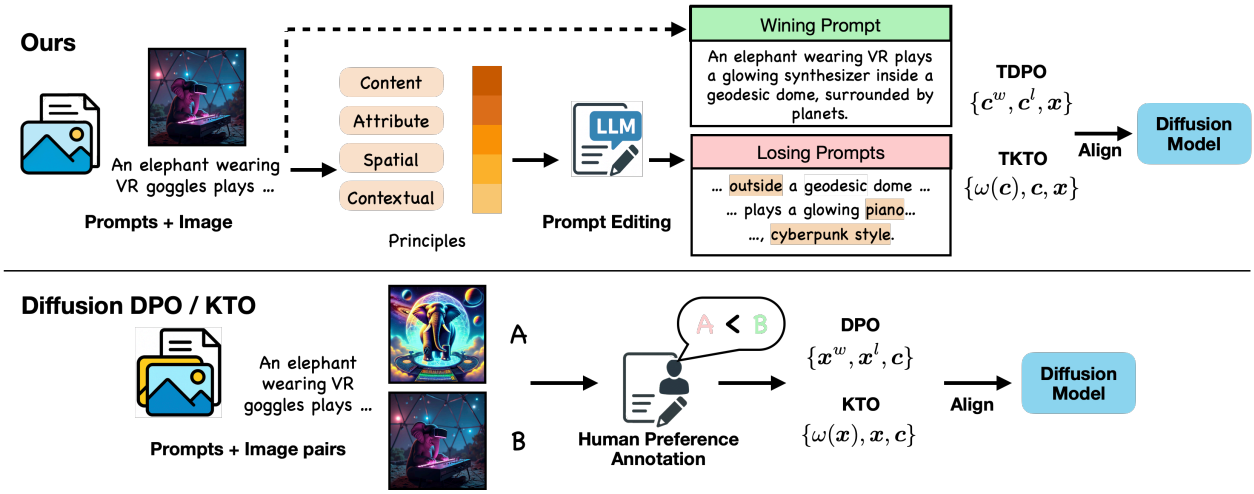

Figure 1: Overview of our Text Preference Optimization (TPO) alignment framework versus the standard Diffusion-DPO/KTO pipeline. **(Top)** We leverage LLMs to perform prompt editing under four principles (content, attribute, spatial, contextual), automatically generating mismatched prompts to form winning/losing text pairs. These prompt pairs are then used to align the diffusion model via our TDPO and TKTO variants in a "free lunch" manner. **(Bottom)** In contrast, existing Diffusion-DPO/KTO methods rely on costly human-annotated image preference pairs.

smoothed preference distribution. DSPO Zhu et al. (2025) aims to close the gap between preference-alignment methods in LLMs and T2I diffusion models using score matching. DiffsuionRPO Gu et al. (2024) also use prompt-image mismatch to improve preference alignment. However, their contrast is still between preferred and rejected images, and preference image annotations remain required. TPO instead starts from single image-caption pairs and constructs text-side preferences (c^+, c^-) for the same image, changing the *source* of supervision. The two are orthogonal to each other and can be potentially combined. We introduce a unified framework that is agnostic to the specific preference-alignment loss and can embed any diffusion T2I alignment objective. To illustrate its flexibility, we show that both Diffusion-DPO and Diffusion-KTO naturally instantiate within our formulation.

3 Method

The effectiveness of preference alignment methods for diffusion models (Wallace et al., 2024; Li et al., 2024) depends heavily on access to high-quality image preference datasets. However, collecting such datasets at scale is challenging due to several key limitations. First, collecting human preferences is expensive, as it requires substantial manual effort for both annotation and validation. Moreover, when the underlying diffusion model changes (*e.g.*, from Stable Diffusion 1.5 to 3.0), previously collected preference data may no longer be effective, requiring a fresh round of data collection. Second, human preferences reflect a mixture of factors, including image quality, alignment with the text prompt, and subjective aesthetic judgment, making them inherently noisy and inconsistent.

Our motivation stems from the observation that generating matched and mismatched text pairs is significantly easier than collecting image preference pairs. By leveraging LLMs to generate text preference pairs for each image, we can align text-to-image models with minimal human supervision, offering a highly cost-effective alternative to manual preference annotation. This enables scalable and efficient alignment at virtually no additional labeling cost. More formally, given a paired dataset $\{\boldsymbol{x}, \boldsymbol{c}^w, \boldsymbol{c}^l\}$ where $\boldsymbol{x}$ is a high-quality image from a human-curated dataset, and $\boldsymbol{c}^w$ and $\boldsymbol{c}^l$ are the matched and mismatched captions, respectively, we aim to learn a new model $p_\theta(\boldsymbol{x}|\boldsymbol{c})$ that achieves better text-to-image alignment and human preference alignment. Below, we first introduce our text preference pair construction pipeline and then formally derive our method.

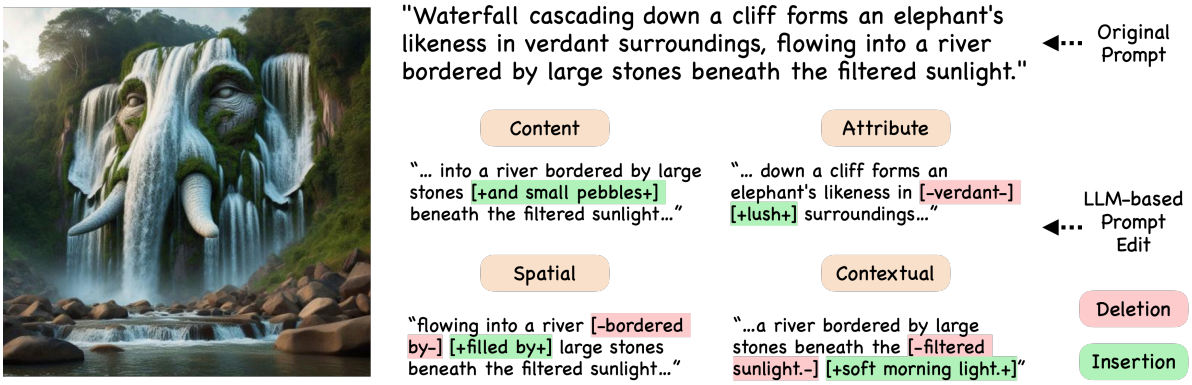

Figure 2: An example of how the four modification principles (content, attribute, spatial, contextual) are applied on a given image-prompt pair.

3.1 Text preference construction with LLMs

Given a paired dataset $\{\boldsymbol{x}, \boldsymbol{c}\}$, where $\boldsymbol{x}$ is an image and $\boldsymbol{c}$ is its corresponding caption, we aim to construct a new dataset with triplets $\boldsymbol{x}, \boldsymbol{c}^w, \boldsymbol{c}^l$, where $\boldsymbol{c}^w$ aligns better with the image $\boldsymbol{x}$ than $\boldsymbol{c}^l$. We assume the original dataset is of high quality, meaning that the caption $\boldsymbol{c}$ accurately describes the image. Based on this assumption, we use LLMs to generate negative captions $\boldsymbol{c}^l$ while using the original caption from the dataset as $\boldsymbol{c}^w = \boldsymbol{c}$. To achieve this, we prompt LLMs to modify the ground truth caption such that the perturbed caption describes an image that is visually distinct from the original. More specifically, we define the four core principles for LLMs to follow when editing the prompt:

- **Content Modifications**: Altering the presence or quantity of concrete objects in the scene. This includes adding or removing objects, replacing one object with another, or changing the number of instances (*e.g.*, modifying "three trees" to "five trees").
- **Attribute and Descriptor Modifications**: Changing visual or stylistic properties of objects, such as material, texture, or style. This also includes enriching descriptions with detailed qualifiers, while avoiding trivial edits like simple color changes.
- **Spatial and Dynamic Modifications**: Modifying the spatial arrangement or motion state of objects. Examples include adjusting object poses, changing spatial relations (foreground/background), or altering object alignment and composition.
- **Contextual and Environmental Modifications**: Editing elements related to the broader scene context, such as background, weather, lighting, or time of day. Changes may also involve shifting the physical or cultural setting (*e.g.*, urban vs. historical).

An example of how these four principles are applied is shown in Fig. 2. We provide more examples and details in the supplementary material.

3.2 Preference Optimization over Input Conditions

In our setting, given an image $\boldsymbol{x}$, we have a matching text prompt $\boldsymbol{c}^w$ and a set of mismatched prompts $\{\boldsymbol{c}_i^l\}_{i=1}^N$. Unlike standard DPO and KTO, where preference optimization is performed over output images, our goal is to align the model with both matched and mismatched input conditions. In this section, we derive the text preference optimization objectives for DPO and KTO under our setting, which we refer to as Text Preference DPO (TDPO) and Text Preference KTO (TKTO), respectively.

TDPO. For Text Preference DPO, we have a triplet $\{\boldsymbol{c}^w, \boldsymbol{c}^l, \boldsymbol{x}\}$ for each image. We have access to the preference relation $\boldsymbol{c}^w \succ \boldsymbol{c}^l \mid \boldsymbol{x}$, where $\boldsymbol{c}^w$ and $\boldsymbol{c}^l \in \{\boldsymbol{c}_i^l\}_{i=1}^N$ denote the matched and mismatched text prompts,

respectively. Following Rafailov et al. (2023); Wallace et al. (2024), we use the Bradley-Terry (BT) (Bradley & Terry, 1952) model to model this preference as

$$p_{\mathrm{BT}}(\boldsymbol{c}^w \succ \boldsymbol{c}^l \mid \boldsymbol{x}) = \sigma(r(\boldsymbol{c}^w, \boldsymbol{x}) - r(\boldsymbol{c}^l, \boldsymbol{x})), \tag{1}$$

where σ is the sigmoid function and the reward model r_θ is a neural network parameterized by θ. Our training objective is to minimize the negative log-likelihood of this preference:

$$L = -\mathbb{E}_{\boldsymbol{c}^w, \boldsymbol{c}^l, \boldsymbol{x}}[\log \sigma(r(\boldsymbol{c}^w, \boldsymbol{x}) - r(\boldsymbol{c}^l, \boldsymbol{x}))] \tag{2}$$

To simplify this optimization and avoid explicit reward modeling, DPO implicitly represents the reward function. Following the derivations in Rafailov et al. (2023), the reward function can be expressed by optimal p^*_θ and p_{ref}, and by taking Bayes' rule, we get:

$$r(\boldsymbol{c}, \boldsymbol{x}) = \beta \log \frac{p^*_\theta(\boldsymbol{c}|\boldsymbol{x})}{p_{\mathrm{ref}}(\boldsymbol{c}|\boldsymbol{x})} + \beta \log Z(\boldsymbol{x}) = \beta \left[\log \frac{p^*_\theta(\boldsymbol{x}|\boldsymbol{c})}{p_{\mathrm{ref}}(\boldsymbol{x}|\boldsymbol{c})} - \log \frac{p^*_\theta(\boldsymbol{x})}{p_{\mathrm{ref}}(\boldsymbol{x})}\right] + \beta \log Z(\boldsymbol{x}) \tag{3}$$

In the second equality of Eq. (3), we assume that the text condition is sampled from the dataset $c \sim D_c$ and is independent of the model parameters θ. Consequently, we have $p_\theta(\boldsymbol{c}) = p_{\mathrm{ref}}(\boldsymbol{c})$. However, $p_\theta(\boldsymbol{x}) \neq p_{\mathrm{ref}}(\boldsymbol{x})$ since the distribution of the generated image is characterized by θ. We can then substitute Eq. (3) into Eq. (2) and get:

$$L_{\mathrm{TDPO}}(\theta) = -\mathbb{E}_{\boldsymbol{c}^w, \boldsymbol{c}^l, \boldsymbol{x}}\left[\log \sigma\left(\beta \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c}^w)}{p_{\mathrm{ref}}(\boldsymbol{x}|\boldsymbol{c}^w)} - \beta \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c}^l)}{p_{\mathrm{ref}}(\boldsymbol{x}|\boldsymbol{c}^l)}\right)\right] \tag{4}$$

where optimizing Eq. (4) is equivalent to maximizing Eq. (1).

TKTO. For Text Preference KTO, we have input $\{\boldsymbol{c}, \boldsymbol{x}, \omega(\boldsymbol{c})\}$ where $\omega(\boldsymbol{c}) = 1$ if $\boldsymbol{c} = \boldsymbol{c}^w$ and $\omega(\boldsymbol{c}) = -1$ for $\boldsymbol{c} \in \{\boldsymbol{c}^l_i\}_{i=1}^N$. KTO seeks to maximize the following objective:

$$\max_r \quad \mathbb{E}_{\boldsymbol{c}, \boldsymbol{x}}[U(\omega(\boldsymbol{c})\,(r(\boldsymbol{c}, \boldsymbol{x}) - z_0))] \tag{5}$$

Here U is the utility function, where $r(\boldsymbol{c}, \boldsymbol{x})$ and z_0 are defined as:

$$r(\boldsymbol{c}, \boldsymbol{x}) = \beta \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c})}{p_{\mathrm{ref}}(\boldsymbol{x}|\boldsymbol{c})}; \quad z_0 = \mathrm{sg}\left[\beta KL(p_\theta(\boldsymbol{x}|\boldsymbol{c})||p_{\mathrm{ref}}(\boldsymbol{x}|\boldsymbol{c}))\right] \tag{6}$$

where sg refers to the stop gradient operator. Following Kahneman-Tversky's prospect theory (Tversky & Kahneman, 1992) , we use a centered sigmoid function as utility function. This gives us the training objective for TKTO:

$$L_{\mathrm{TKTO}}(\theta) = -\mathbb{E}_{\boldsymbol{c}, \boldsymbol{x}}\left[-\sigma(\omega(\boldsymbol{c})\beta(\log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c})}{p_{\mathrm{ref}}(\boldsymbol{x}|\boldsymbol{c})} - z_0))\right] \tag{7}$$

3.3 Text Preference Optimization for Diffusion Model

We now extend our TPO alignment algorithm to align diffusion-based T2I models.

Diffusion Model. Diffusion models (Song et al., 2020; Ho et al., 2020; Kingma et al., 2023) are generative models that sample from a learned distribution $p_\theta(\boldsymbol{x}_0)$, trained to approximate the empirical data distribution $q(\boldsymbol{x}_0)$. Training proceeds by learning to invert a fixed forward process of (denoising) diffusion $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. The forward process $q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)$ is a Markov chain with Gaussian transition probabilities governed by noise schedules $\{\alpha_t, \sigma_t\}$, as defined in Rombach et al. (2022a), that gradually add noise to data $\boldsymbol{x}_0$. The reverse (denoising) process is defined by

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta\big(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\big), \quad \text{where } p_\theta\big(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\big) = \mathcal{N}\Big(\boldsymbol{x}_{t-1};\, \mu_\theta(\boldsymbol{x}_t, t),\ \tilde{\sigma}_t^2\, I\Big)\,. \tag{8}$$

A neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is trained to predict the noise ϵ in $\boldsymbol{x}_t$ by minimizing the simplified evidence lower bound associated to this model,

$$L_{\mathrm{DM}} = \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}),\, t \sim \mathcal{U}[0,T],\, \boldsymbol{\epsilon} \sim \mathcal{N}(0,I), \boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\Big[w(t)\, \big\|\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\big\|_2^2\Big], \tag{9}$$

where $w(t)$ is a weighting function.

Diffusion TDPO. Existing work has adapted DPO to diffusion models (Wallace et al., 2024; Yang et al., 2024). Diffusion-DPO (Wallace et al., 2024) frames the diffusion process as an MDP and defines the reward as the reward of the whole chain, $r(\boldsymbol{c}, \boldsymbol{x}_0) = \mathbb{E}_{p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0,\boldsymbol{c})}[R(\boldsymbol{c}, \boldsymbol{x}_{0:T})]$. It then utilizes an evidence lower bound to adapt the DPO training objective to the diffusion-model setting. For details of this adaptation, please see Section 4 of Diffusion-DPO (Wallace et al., 2024). Following their work and Eq. (4), we can similarly derive the naive DiffusionTDPO objective, it can be rewritten as:

$$\begin{aligned} L_{\text{Diff-TDPO}} = -\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{c}^w,\boldsymbol{c}^l,t,\boldsymbol{x}_t}[\log \sigma(-\beta w(t)(||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}^w, t) - \boldsymbol{\epsilon}^w||_2^2 - ||\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t, \boldsymbol{c}^w, t) - \boldsymbol{\epsilon}^w||_2^2 \\ - (||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}^l, t) - \boldsymbol{\epsilon}^l||_2^2 - ||\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t, \boldsymbol{c}^l, t) - \boldsymbol{\epsilon}^l||_2^2)))], \end{aligned} \tag{10}$$

where $\boldsymbol{\epsilon}_\theta$ and $\boldsymbol{\epsilon}_{\text{ref}}$ are our training model and pretrained frozen model, $\mathcal{U}$ is a uniform distribution.

Diffusion TKTO. Following previous work on adapting DPO to diffusion models, another alignment optimization algorithm, KTO, has also been adapted to this field. Based on Diffusion-KTO (Li et al., 2024) and Eq. (11), we present Diffusion-TKTO as:

$$L_{\text{Diff-TKTO}} = -\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{c},t,\boldsymbol{x}_t}\sigma(w(\boldsymbol{c})\beta\left[-(\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})\|_2^2 - \|\epsilon - \epsilon_{\text{ref}}(\boldsymbol{x}_t, t, \boldsymbol{c})\|_2^2) - z_0\right]) \tag{11}$$

here $w(\boldsymbol{c}) = \pm 1$ if the prompt $\boldsymbol{c}_0$ is matched or mismatched. Consistent with Li et al. (2024), we set $z_0 = \text{sg}\left[\beta KL(p_\theta(\boldsymbol{x}|\boldsymbol{c})||p_{\text{ref}}(\boldsymbol{x}|\boldsymbol{c}))\right]$.In practice, we use a biased but low-variance estimator for z_0:

$$\max\left(0,\ \frac{1}{m}\sum \beta(-\left(\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})\|_2^2 - \|\epsilon - \epsilon_{\text{ref}}(\boldsymbol{x}_t, t, \boldsymbol{c})\|_2^2\right))\right)$$

4 Experiments

4.1 Experiments Setup

Datasets and Model. In this work, all experiments are conducted using Stable Diffusion v1.5 (SD v1.5) (Rombach et al., 2022b), SDXL Podell et al. (2023) and Stable Diffusion 3.5 Medium (SD35M) Esser et al. (2024b). We fine-tune SD v1.5 on the Human Preference Synthetic Dataset (HPSD) (Egan et al., 2024) and Pick-a-Pic v2 Kirstain et al. (2023). We also finetune SDXL on Pick-a-Pic v2. Following previous work (Wallace et al., 2024; Li et al., 2024), we use HPDv2 test set (Wu et al., 2023b), Pick-a-Pic v2 test dataset (Kirstain et al., 2023) and Parti-Prompts dataset (Yu et al., 2022) for evaluation. In addition, we also include the open-image-preferences-v1 (Berenstein et al., 2024) dataset for evaluation. This dataset contains prompts collected from everyday image-generation user requests from online platforms and has been filtered through a toxicity-reduction pipeline, making it an especially robust evaluation suite. Details on each dataset are provided in supplementary material.

Training Setup. Our training of SD 1.5 involves two stages, both of which fine-tune only the U-Net of SD v1.5 while freezing all other components. In the first stage, we fine-tune the pretrained SD v1.5 model on the training dataset using a pure SFT loss objective until convergence is observed. This stage is to close the gap between the pretrained model and the target dataset, such SFT stage is common in RLHF fine-tuning in LLM literature (Christiano et al., 2023; Ouyang et al., 2022). In Stage 2, we continue fine-tuning the SFT-adapted model with identical hyperparameters using our TDPO and TKTO methods, alongside the Diffusion-DPO and Diffusion-KTO baselines. We also experimented on SDXL and SD35M, as SFT training on these models would only degrade model performance. Our SDXL and SD35M experiment starts with 1-stage TDPO and TKTO training. In our experiments, we observed that direct fine-tuning with the objectives in Eq. (10) and Eq. (11) degraded sample quality after several steps. To address this, we introduce a clipping mechanism inspired by Proximal Policy Optimization (Schulman et al., 2017) to improve training stability. Concretely, we clamp the squared L2 norm of the negative-sample term in the loss, thereby bounding extreme negative signals and stabilizing training (see the supplementary material for details).

Evaluation. In our evaluation, we generate images from each model using the same test prompts and assess alignment with four metrics: PickScore (Kirstain et al., 2023), CLIP alignment (Radford et al., 2021), HPSv2

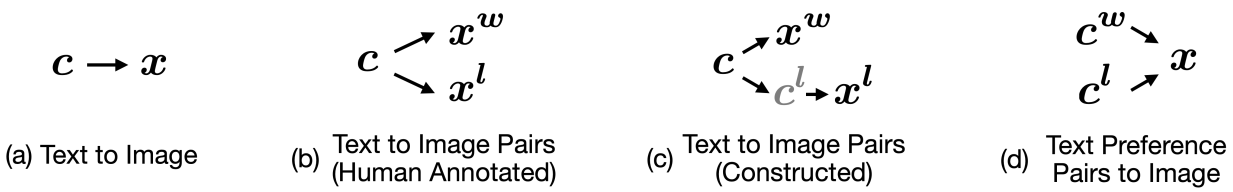

Figure 3: Comparison of setups for different text-to-image alignment methods. (a) Supervised fine-tuning; (b,c) Diffusion-KTO, Diffusion-DPO, DSPO, MaPO, etc.; (d) Ours.

(Wu et al., 2023b), and ImageReward (Xu et al., 2023). For each metric, we compute the win rate against the SD v1.5 (or SDXL) baseline, i.e., the fraction of prompts on which a model's score exceeds SD v1.5's (or SDXL's), and report the average win rate across all four metrics. All evaluations use identical sampling settings (fixed random seed, classifier-free guidance scale of 7.5, and 50 diffusion steps).

4.2 Comparison Protocol and Evaluation Settings

In Fig. 3, we organize text-to-image alignment methods into four training configurations:

- (a) fine-tuning with desired prompt–image pairs;
- (b) alignment with human-annotated preferred/dispreferred image pairs;
- (c) synthetic image-pair alignment by editing prompts and generating less-preferred images;
- (d) our approach, which aligns with positive/negative prompt pairs without requiring negative images.

This protocol is designed to answer two questions for DiffusionTPO: (1) when human preference pairs are available, how close can text-pair supervision get to methods trained with oracle human preferences? (2) when human preference pairs are unavailable, how much improvement can DiffusionTPO deliver over existing alternatives?

Accordingly, we evaluate all methods in two settings: (1) **Human-annotated preference pairs available** (setting (b) in Fig. 3); (2) **Only text-to-single-image data available** (settings (c)/(d), with no human-annotated preference pairs). The first setting follows prior work and serves as a reference setting. The second is our target scenario; the next two subsections report results in this order.

4.3 Quantitative Results with Human-Annotated Preference Pairs

Tab. 1 We first report setting (b), where human preference annotations are available through Pick-a-Pic v2 (Kirstain et al., 2023). In this setting, competing methods are trained on annotated preferred/dispreferred image pairs, while our method uses only half the dataset, with captions and winning images but without preference-pair supervision. As shown in Tab. 1, TDPO consistently outperforms Diffusion-DPO. TKTO is slightly below Diffusion-KTO on HPS and ImageReward. Nevertheless, our approach remains competitive while requiring no human preference annotation. These results provide a strong reference point, validating our method's potential for aligning T2I models relative to approaches that require oracle human preferences.

4.4 Quantitative Results without Human Preference Pairs

We now turn to the setting where no human-annotated preference pairs are available. Tab. 2 reports win rates against SD v1.5 across four datasets and four evaluation metrics. For image-pair baselines (Diffusion-DPO and Diffusion-KTO), we keep the positive image fixed as the original high-quality dataset image x^w and only construct the negative image x^l. We construct x^l with an off-the-shelf text-to-image model conditioned on our edited negative prompt c^l, using two variants: $x^l = \text{SD3.5}(c^l)$ and $x^l = \text{SD1.5-FT}(c^l)$, where SD1.5-FT denotes SD v1.5 fine-tuned on the current text-to-image training dataset. Our method uses the same negative prompts c^l but does not require generating negative images (text-only supervision). This design makes the

Table 1: Comparison to previous methods trained on the pick-a-pick human preference dataset with Win rates (%) against SD-1.5 are reported. Our method only uses the winning image x^w. Metrics with * are directly taken from Zhu et al. (2025).

Method	**Supervision**	Pick-a-Pic v2				Parti-Prompt			
		PS	**CLIP**	**HPS**	**IR**	**PS**	**CLIP**	**HPS**	**IR**
SFT	(x, c)	70.80	54.80	79.40	76.80	67.38	54.31	75.00	67.69
Diff.-DPO	(x^w, x^l, c)	72.80	57.80	69.20	65.20	66.88	57.31	65.63	61.50
Ours-TDPO	(x, c^w, c^l)	70.60	62.60	75.80	75.80	67.12	57.06	69.88	67.75
Diff.-KTO	(x^w, x^l, c)	74.60	54.80	80.60	77.60	63.75	55.00	75.38	69.19
Ours-TKTO	(x, c^w, c^l)	**74.60**	**63.80**	71.40	74.00	65.25	59.63	68.88	67.75
DSPO*	(x^w, x^l, c)	73.60	61.80	**84.80**	**78.00**	65.32	54.86	**87.50**	71.75
InPO	(x^w, x^l, c)	72.81	60.19	78.69	74.94	**83.20**	61.00	79.20	**82.00**
SmPO	(x^w, x^l, c)	73.06	61.81	77.94	74.56	83.00	**64.20**	79.20	79.40

Table 2: Comparison against Diffusion-DPO and Diffusion-KTO baselines in the text-to-single-image setting. Win rates (%) against SD v1.5 are reported. Best in **bold**, second in underline. Our methods are shaded in blue. The column "x^l source" specifies how negative images are generated for pair-based baselines.

		HPSv2				PartyPrompt			
Method	**x^l source**	**PS**	**CLIP**	**HPS**	**IR**	**PS**	**CLIP**	**HPS**	**IR**
SFT	–	76.25	52.00	75.75	76.00	69.06	52.38	63.31	73.81
Diff-DPO-sd35	$x^l = \text{SD3.5}(c^l)$	79.50	49.00	76.50	74.50	71.94	48.69	64.06	73.37
Diff-DPO-SFTsd15	$x^l = \text{SD1.5-FT}(c^l)$	77.00	54.75	59.00	70.00	71.06	54.06	49.94	64.88
Ours-TDPO	text-only (c^l)	**83.25**	**56.00**	79.00	**82.25**	**74.63**	57.25	70.19	**78.75**
Diff-KTO-sd35	$x^l = \text{SD3.5}(c^l)$	77.25	52.25	72.25	75.00	68.06	50.94	63.69	70.81
Diff-KTO-SFTsd15	$x^l = \text{SD1.5-FT}(c^l)$	80.25	53.75	76.00	76.75	73.50	52.63	63.69	71.62
Ours-TKTO	text-only (c^l)	80.00	55.75	**81.00**	80.75	73.62	**57.31**	**70.63**	77.31
		Pick-a-Pic v2				OpenImagePref			
SFT	–	71.20	50.80	62.00	72.80	80.60	53.60	77.00	77.60
Diff-DPO-sd35	$x^l = \text{SD3.5}(c^l)$	72.40	52.20	66.00	76.80	81.40	46.60	73.60	77.40
Diff-DPO-SFTsd15	$x^l = \text{SD1.5-FT}(c^l)$	72.80	54.20	52.40	68.60	83.60	57.20	62.00	73.20
Ours-TDPO	text-only (c^l)	75.20	56.00	69.20	**82.60**	**88.20**	58.20	**80.20**	**86.00**
Diff-KTO-sd35	$x^l = \text{SD3.5}(c^l)$	72.20	54.00	64.40	74.60	77.80	54.60	75.80	77.00
Diff-KTO-SFTsd15	$x^l = \text{SD1.5-FT}(c^l)$	**77.40**	55.20	65.60	74.60	82.60	57.00	76.60	80.80
Ours-TKTO	text-only (c^l)	74.40	**58.20**	**70.80**	80.20	84.20	**60.80**	79.20	82.40

comparison explicit and fair: all methods share the same fixed positive target x^w and differ only in the training objective and in the generator used to construct x^l for image-pair baselines. For example, in Tab. 2, the value 83.25 (Ours-TDPO, PickScore, HPSv2) means that TDPO outperforms SD v1.5 on 83.25% of prompts under that metric. Overall, our methods outperform most baselines, with Diffusion-KTO remaining slightly higher on a few individual metrics. More ablations and baseline details are provided in the supplementary material.

4.5 Qualitative Results and Human Study

After reporting quantitative results in both settings, we now provide qualitative comparisons and human-preference validation.

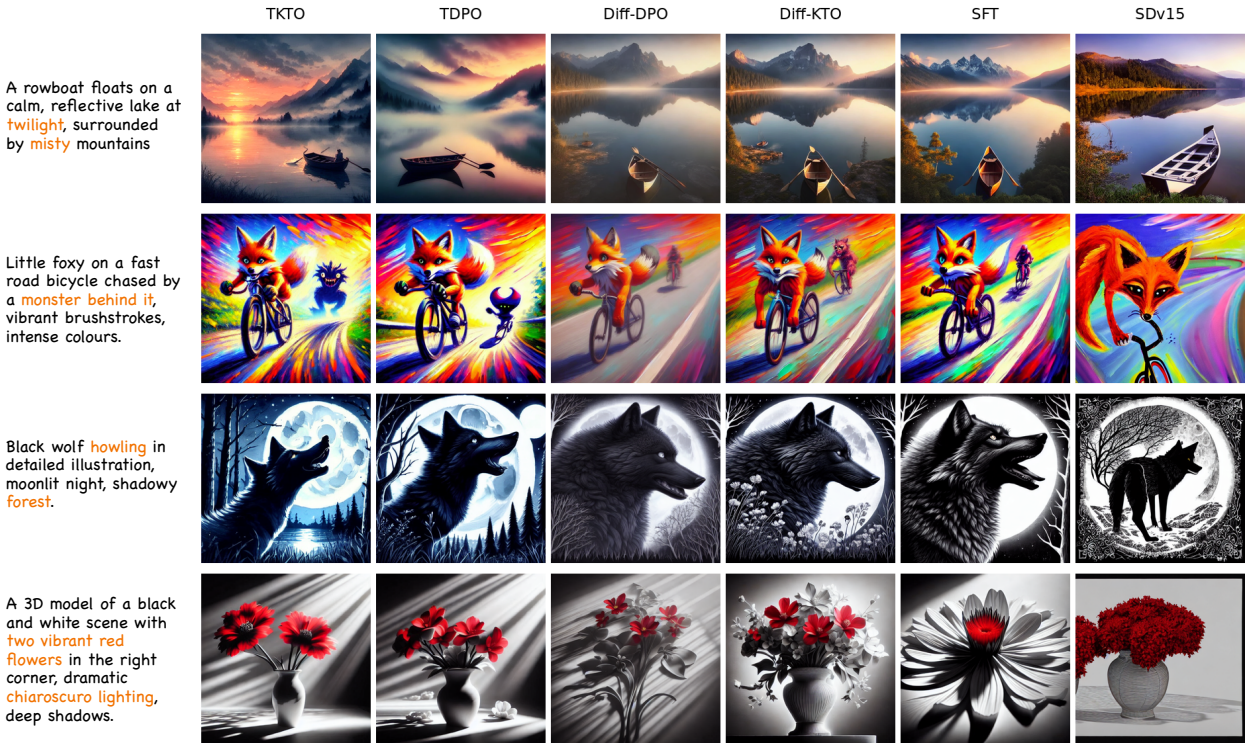

Figure 4: Side-by-Side grid comparison of the image generation using our methods and the baselines. The leftmost column is the prompts used to generate the images. The important concept or element of the prompt that the model successfully or fails to capture is printed in orange color.

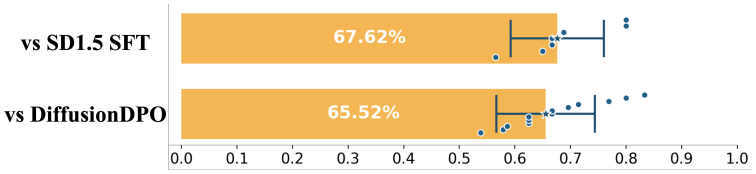

Figure 5: Human studies on prompt-to-image alignment. Top: SD1.5 TDPO vs SD1.5 SFT. Bottom: SD1.5 TDPO vs Diffusion-DPO SD1.5.

Qualitative Results. In Fig. 4, we show a side-by-side comparison of images generated by our method and by the baselines. In the first row, we can see that only our methods, TDPO and TKTO, render both "twilight" and "misty" distinctly. Other baselines struggle to convey either concept. In the second row, all methods except vanilla SD v1.5 hint at something "behind" the biker, but only ours clearly shows a monster in pursuit. The others look more like a second rider tailing the first. In the third row, our methods and SFT successfully generate a black wolf that is "howling", and only our methods also embed the "forest" setting. The other methods either omit the trees altogether or render them too faintly. In the last row, we can see that the TKTO method precisely generates two red flowers against a dark background with strong, directional light and deep shadows. The baselines produce the wrong number of red flowers and offer only weak chiaroscuro effects. Overall, these examples demonstrate that TKTO consistently captures each prompt's detailed concepts and produces images that align more faithfully with what the user asked for.

Human Study. To complement the automatic metrics, we conduct a human study comparing Diffusion-DPO, supervised fine-tuning (SFT), and our TDPO method on prompt–image alignment. For Diffusion-DPO, we use the original checkpoint provided in Wallace et al. (2024), trained on Pick-a-Pic v2 (Kirstain et al., 2023). For TDPO, we adopt the same setting as Tab. 1, where only winning images from Pick-a-Pic v2 are used. For the SFT baseline, we fine-tune SD1.5 on the same winning image–prompt pairs. This setting keeps

Table 3: Alignment Performance comparison of SDXL. Win rates (%) are reported against SDXL. Note that our method only requires a single image x while others require a pair.

Method	Supervision	Pick-a-Pic v2				Parti-Prompt			
		PS	**CLIP**	**HPS**	**IR**	**PS**	**CLIP**	**HPS**	**IR**
SFT*	(x, c)	20.80	44.80	40.60	34.40	17.03	36.58	33.02	37.18
Diff.-DPO	(x^w, x^l, c)	64.40	60.80	62.40	63.60	61.19	58.13	70.69	71.25
MaPO	(x^w, x^l, c)	45.60	47.60	58.60	54.60	44.31	41.38	56.25	58.00
InPO	(x^w, x^l, c)	67.20	52.80	69.40	62.80	66.63	49.63	69.63	66.63
SmPO	(x^w, x^l, c)	**80.80**	55.20	**75.80**	**73.60**	**82.69**	54.94	**81.38**	**77.38**
Ours-TDPO	(x, c^w, c^l)	70.60	**63.60**	72.60	66.80	**70.43**	59.69	69.82	70.18
Ours-TKTO	(x, c^w, c^l)	72.20	63.40	71.80	68.20	68.65	**59.82**	67.91	68.10

Method	Supervision	Pick-a-Pic v2				Parti-Prompt			
		PS	**CLIP**	**HPS**	**IR**	**PS**	**CLIP**	**HPS**	**IR**
Diff.-DPO	(x^w, x^l, c)	51.00	48.40	53.20	49.60	52.23	50.55	48.28	52.21
Ours-TDPO	(x, c^w, c^l)	**58.20**	**58.00**	**63.80**	**68.80**	**58.71**	**55.83**	**64.85**	**64.36**
Diff.-KTO	(x^w, x^l, c)	**63.80**	53.40	58.00	64.40	**60.74**	51.10	55.64	60.61
Ours-TKTO	(x, c^w, c^l)	51.40	**56.80**	**67.60**	**69.60**	53.37	**54.00**	**61.17**	**65.03**

Table 4: Stable Diffusion 3.5 Medium (SD35M) preference alignment. **Bold** marks the winner within each base method (DPO or KTO).

the comparison fair: all methods are trained on the same winning pairs, while Diffusion-DPO additionally uses human-dispreferred images and TDPO instead leverages generated negative prompts. More details on human-study collection are provided in the supplementary material.

4.6 Generalization of TPO

Experiment beyond SD 1.5 To validate that our method works beyond SD1.5, we provide additional results on the SDXL base model in Tab. 3. We adopt the same evaluation configuration as Tab. 1: competing methods are trained using both preferred and dispreferred images from Pick-a-Pic v2, whereas our method relies solely on the winning image. The results indicate that our method achieves superior CLIP scores, while performing comparably for other scores. Importantly, it provides a clear improvement over the SDXL base model and surpasses the SFT baseline, whose performance degrades due to the lower visual quality of Pick-a-Pic v2 images relative to SDXL's base outputs.

Generalization to Flow matching models. Diffusion model and Flow matching models are similar in nature, here we evaluate TPO on SD3.5M, which adopts a flow-matching objective and an MM-DiT architecture – substantially different from SD-1.5/SD-XL series. Tab. 4 shows TDPO/TKTO consistently improve over Diff.-DPO/KTO, indicating that TPO generalizes beyond standard diffusion U-Nets.

Compatibility with image-pair preference methods. We aim to address below two questions: Can TPO further improve a model that is already image-pair preference-aligned? Tab. 5 answers *yes* to it. In the table, *Stage 1=Diff.-DPO/SmPO, Stage 2=TDPO/TKTO* addresses the question, testing TPO as a *post*-alignment step on top of an image-pair-aligned model on SDXL. Values are deltas vs. the block baseline (in *italic*); cells use green/red for gain/loss, and bold marks the per-column best in each block. We find that TPO can continue to improve image-pair preference-aligned models when applied as stage 2. This suggests that improvement brought by image-pair PO and TPO are largely complementary to each other, where TPO layers cleanly on existing pipelines as a free, plug-and-play step.

Stage 1	Stage 2	Pick-a-Pic v2				Parti-Prompt			
		PS	CLIP	HPS	IR	PS	CLIP	HPS	IR
Diff.-DPO	–	*64.40*	*60.80*	*62.40*	*63.60*	*61.19*	*58.13*	*70.69*	*71.25*
Diff.-DPO	TDPO	+6.80	+1.40	+2.80	+4.60	+5.07	+3.65	-8.85	+0.71
Diff.-DPO	TKTO	**+8.00**	+2.80	-0.40	+7.80	**+6.97**	**+4.81**	-6.95	+1.76
SmPO	–	*80.80*	*55.20*	*75.80*	*73.60*	***82.69***	*54.94*	***81.38***	*77.38*
SmPO	TDPO	-0.40	+7.20	+2.00	-1.80	-1.83	+7.27	-5.00	**+0.17**
SmPO	TKTO	**+0.40**	**+9.60**	**+2.20**	**+1.00**	-4.29	**+7.82**	-4.26	+0.10

Table 5: TPO combined with image-pair PO on SDXL.

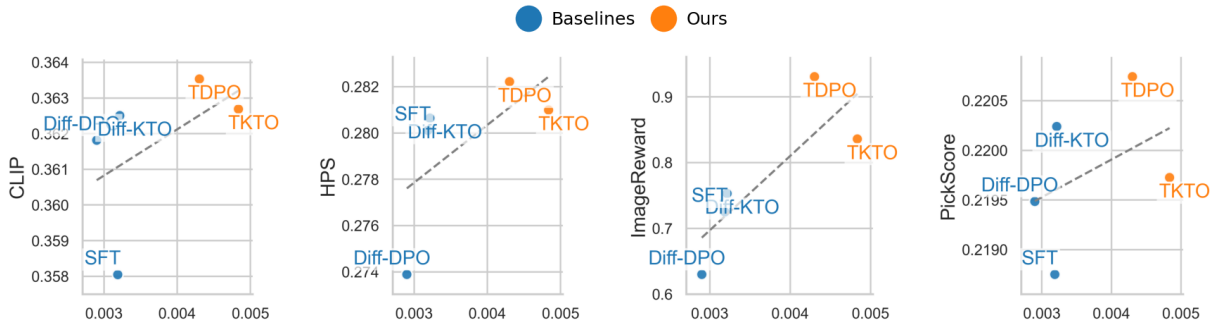

Figure 6: Regression plots between alignment metrics and implicit preference score. These plots show a positive correlation between alignment metrics and implicit preference score. Our method consistently achieves higher implicit preference scores and higher human preference scores.

5 Ablations and Discussions

Implicit preference over text prompts correlates with human preferences over images. We aim to answer the following question: *How does learning from preferences over text prompts improve human preferences?* Here we shed light on this matter in Fig. 6. To begin with, we define a metric called the implicit preference score (IPS), which quantifies how much a model prefers $\boldsymbol{x}, \boldsymbol{c}^w$ over $\boldsymbol{x}, \boldsymbol{c}^l$ as follows:

$$\textbf{IPS} = \mathbb{E}_{t\sim\mathcal{U},\boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \boldsymbol{c}^l)\|_2^2 - \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \boldsymbol{c}^w)\|_2^2]. \tag{12}$$

The term, intuitively interpretable as the difference in diffusion loss between negative and positive pairs, quantifies how much more likely the model is to generate image $\boldsymbol{x}$ given the matching prompt $\boldsymbol{c}^w$ compared to the mismatched prompt $\boldsymbol{c}^l$. A larger value indicates better alignment with the ground truth textual preference. The regression plots in Fig. 6 reveal a clear positive correlation between human preference metrics and the implicit preference score: models with higher preference scores exhibit higher human preference metrics. *This suggests an underlying connection between alignment with textual prompt pairs and alignment with human image preferences.* Our methods, including their TDPO and TKTO variants, appear in the top-right region of the plots, indicating both the highest implicit preference score and the highest human preference scores among almost all evaluated baselines.

The effectiveness of prompt editing principles. Despite the excellent performance improvement shown in Tab. 2, we still need to answer the question: *What benefit does each editing principle provide?* To this end, we conduct an ablation study on how each editing principle defined in Sec. 3.1 affects alignment performance. Specifically, we fine-tune under four different settings; in each setting, one editing principle is dropped. Tab. 6 shows that removing different modification principles leads to different performance across metrics and methods. Interestingly, we observe that removing content-related modifications leads to a significant drop in CLIP score. This is likely because content modifications most directly influence the model's sensitivity to the semantic meaning of prompts. We also observe that dropping spatial-related modifications leads to a

Table 6: Ablation on prompt editing principles. Left: TDPO. Right: TKTO. Metrics are reported as win rates (%).

TDPO					TKTO				
Model	PS	CLIP	HPS	IR	Model	PS	CLIP	HPS	IR
TDPO	74.6	57.3	70.2	78.8	TKTO	73.6	57.3	70.6	77.3
w/o Attribute	74.3	56.3	71.6	78.3	w/o Attribute	72.4	57.4	70.1	78.2
w/o Content	75.2	55.6	67.4	76.1	w/o Content	74.9	56.1	69	76.7
w/o Contextual	73.6	57.1	70.8	77.0	w/o Contextual	69.2	56.1	67.6	73.8
w/o Spatial	76.4	56.4	70.4	77.2	w/o Spatial	74.4	57	72.8	78

Table 7: Ablation on rewriting LLM's effect on the final performance.

Method	**Rewriting LLM**	Pick-a-Pic v2				Parti-Prompt			
		PS	**CLIP**	**HPS**	**IR**	**PS**	**CLIP**	**HPS**	**IR**
SFT	–	71.20	50.80	62.00	72.80	69.06	52.38	63.31	73.81
TDPO	Gemini-Flash-2.0	75.20	56.00	69.20	82.60	74.63	57.25	70.19	78.75
	Qwen3-30b-a3b-it	75.20	54.80	64.40	78.40	72.19	54.94	65.19	76.25
	Llama-3.1-8b-it	68.20	54.00	61.00	72.80	67.44	54.00	64.44	75.81

trade-off, and even improvement for TKTO. A plausible cause is that spatial prompts inject high-variance, low-signal supervision: the viewpoint of the image is often underspecified, which brings ambiguity to spatial predicates like "left" and "right", and enforcing spatial changes entangles cross-attention with object layout, which destabilizes overall learning.

The effect of large language model choice for rewriting Lastly, we examine the effect of choosing a large language model for prompt rewriting. Specifically, we evaluate the open-source alternatives Qwen3-30B-A3B-it and Llama-3.1-8b-it. We show the performance comparison of these three models for generating negative prompts in Tab. 7. It can be seen that Gemini-Flash-2.0 produces the best performance, while Llama-3.1-8b-it fails to provide meaningful improvements over the supervised fine-tuning baseline. This suggests that the general capability of the LLM has a non-negligible impact on the final alignment performance. Notably, open-source alternatives such as Qwen3-30B-A3B-it are sufficiently powerful to deliver improvements over the baseline methods.

6 Conclusion

We have presented a novel "free lunch" alignment method that leverages LLM-generated text–preference pairs to fine-tune text-to-image diffusion models without requiring human preference annotations. Our instantiations, TDPO and TKTO, achieve consistent improvements over the baselines, while remaining model-agnostic and easily integrated into any RLHF-style pipeline. Future work includes extending this framework by integrating other preference-optimization algorithms, applying it to other modalities such as text-to-video and text-to-3D generation, and exploring richer negative-sample generation techniques for greater diversity.

References

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2022.

David Berenstein, Ben Burtenshaw, Daniel Vila, Daniel van Strien, Sayak Paul, Ame Vi, and Linoy Tsaban. Open preference dataset for text-to-image generation, 2024. URL https://huggingface.co/blog/image-preferences.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

Ralph A. Bradley and Milton M. Terry. Rank analysis of incomplete block designs. i. the method of paired comparisons. *Journal of the American Statistical Association*, 48(262):495–507, 1952.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers, 2023. URL https://arxiv.org/abs/2301.00704.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vmSEVL19f.

Ben Egan, Alex Redden, XWAVE, and SilentAntagonist. Dalle3 1 Million+ High Quality Captions, May 2024. URL https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024a.

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, 2024b. URL https://arxiv.org/abs/2403.03206.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023.

Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization, 2024. URL https://arxiv.org/abs/2406.06382.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems*, 37:76177–76209, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Shyamgopal Karthik, Huseyin Coskun, Zeynep Akata, Sergey Tulyakov, Jian Ren, and Anil Kag. Scalable ranked preference optimization for text-to-image generation, 2024. URL https://arxiv.org/abs/2410.18013.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023. URL https://arxiv.org/abs/2107.00630.

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, 2023. URL https://arxiv.org/abs/2305.01569.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image captions are natural prompts for text-to-image models. *arXiv preprint arXiv:2307.08526*, 2023.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024a.

Mushui Liu, Yuhang Ma, Xinfeng Zhang, Yang Zhen, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. *arXiv preprint arXiv:2407.00737*, 2024b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. Inpo: Inversion preference optimization with reparametrized ddim for efficient diffusion model alignment, 2025a. URL https://arxiv.org/abs/2503.18454.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xiaoyin Xu, and Min Zhang. Smoothed preference optimization via renoise inversion for aligning diffusion models with varied human preferences, 2025b. URL https://arxiv.org/abs/2506.02698.

Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. In *Thirty-eighth Conference on Neural Information Processing Systems 2024*. Neural Information Processing Systems Foundation, 2024.

Yanting Miao, William Loh, Suraj Kothawade, Pascal Poupart, Abdullah Rashwan, and Yeqing Li. Subject-driven text-to-image generation via preference-based reinforcement learning, 2024. URL https://arxiv.org/abs/2407.12164.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

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, 2023. URL https://arxiv.org/abs/2307.01952.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2024. URL https://arxiv.org/abs/2310.03739.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tPNHOoZFl9.

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, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022b. URL https://arxiv.org/abs/2112.10752.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, et al. Sg-adapter: Enhancing text-to-image generation with scene graph guidance. *arXiv preprint arXiv:2405.15321*, 2024.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020. URL https://arxiv.org/abs/2006.11807.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323, 1992.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model. *arXiv preprint arXiv:2311.14284*, 2023a.

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, 2023b. URL https://arxiv.org/abs/2306.09341.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment, 2024. URL https://arxiv.org/abs/2405.00675.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model, 2024. URL https://arxiv.org/abs/2311.13231.

Yongjin Yang, Sihyeon Kim, Hojung Jung, Sangmin Bae, SangMook Kim, Se-Young Yun, and Kimin Lee. Automated filtering of human feedback data for aligning text-to-image diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8jvVNPHtVJ.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://api.semanticscholar.org/CorpusID:249926846.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xyfb9HHvMe.

A More Experiment Details

A.1 Implicit preference score

Recall that in Sec. 5 we defined the implicit preference score for a triplet $(\boldsymbol{x}, \boldsymbol{c}^w, \boldsymbol{c}^l)$ as

$$\mathbb{E}_{t\sim\mathcal{U},\boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}^l)\|_2^2 - \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}^w)\|_2^2]. \tag{13}$$

In practice, we fix the diffusion timestep to $t = 0.5$, sample $\boldsymbol{x}_t$ three times for each triplet, and average the resulting diffusion losses to compute implicit preference score. All triplets are drawn from the HPSD evaluation set. The mismatched prompt $\boldsymbol{c}^l$ is generated by applying a single modification to the original prompt, as described in Sec. 5.

We observe a strong negative correlation between implicit preference score and human preference metrics. In other words, models that incur higher diffusion loss on the mismatched prompt $\boldsymbol{c}^l$ relative to the matched prompt $\boldsymbol{c}^w$ consistently receive higher human preference scores. This result aligns with our goal of improving text–image alignment and helps explain why our method achieves superior performance on human preference evaluations.

A.2 Dataset

Here, we list all the dataset we have used in this study, with a short introduction and the usage in this sutdy.

- HPDv2 (Wu et al., 2023b) (Apache license 2.0) is a large-scale (798k preference choices / 430k images), a well-annotated dataset of human preference choices on images generated by text-to-image generative models. We have use the prompt of its test set for evaluation, the test set include 400 data.
- Pick-a-Pic v2 (MIT license) (Kirstain et al., 2023) is a large and open dataset for human feedback in text-to-image generation. We use its test set for evaluation, which contain 500 data.
- Parti-prompts (Yu et al., 2022) (Apache license 2.0 license) is a rich set of over 1600 prompts in English that we release as part of this work. We use the whole dataset for evaluation.
- open-image-preferences-v1 (Berenstein et al., 2024) (Apache license 2.0 license) is a dataset contain over 7k human preference pairs on images generated by powerful text-to-image generative models. All the images pairs are generated by the same prompt, and a preference binary label is provided by human annotators. We split the dataset into train and evaluation set, where the the last 500 data are split for evaluation while the rest form the training set.
- HPSD (Egan et al., 2024) (MIT license) dataset comprises of high-quality AI-generated images sourced from various websites and individuals, it contains over 780k image-prompt paris. For the first stage SFT, we use the whole dataset for training. For the second stage of fine-tuning, we use the first 100K.

A.3 Training Details

All experiments are run on two NVIDIA A100 GPUs using Stable Diffusion v1.5 (CreativeML Open RAIL-M license) and SDXL. For SD v1.5, we use a batch size of 16 and a constant learning rate of 1×10^{-6} in the second stage. For SFT, we employ a batch size of 256, a learning rate of 1×10^{-5}, and train for 17 500 steps until a convergence on training loss has been observed. For SDXL, since SFT on the training set will degrade quality and performance, we apply TPO fine-tuning directly to SDXL. we use a learnig rate of 5×10^{-7}. All models are trained on an image resolution of 512*512. All models are optimized with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$) and a constant learning rate. Our methods (TDPO, TKTO) and the baselines (Diffusion-DPO, Diffusion-KTO) share $\beta = 5000$ for SD v1.5, and use $\beta = 500$ for SDXL. To select the best checkpoint, we sample 200 prompts from the HPSD evaluation set, generate corresponding images, and compute three evaluation metrics. We then choose the checkpoint with the highest score across these metrics for final evaluation.

"An artistic rendition of a feline face featuring metallic and electronic accents; its eyes glow golden and emerald, within an intricate design of circuits and neon tones."

Content

"... circuits and neon tones, [+with a small antenna on its head+]"

Attribute

"... within an [-intricate-] [+cumbersome and complicated+] design of circuits and neon tones."

Spatial

"... within an intricate design of circuits and neon tones, [+looking above. +]"

Contextual

"... golden and emerald, within an intricate design of circuits [-and neon tones-]."

"An image displays a vertical split moon, one side illuminated showing craters & textures, a dark void separates it from a lit spacecraft launching, set against a starry cosmic background."

Content

"... a dark void separates it from a lit [-spacecraft-] [+rocket+] launching, set against .."

Attribute

".. one side illuminated showing [-craters & textures,-] [+craters,+] a dark void separates it from ..."

Spatial

"... a dark void separates it from a lit spacecraft [-launching,-] [+launching slightly above,+] set against .."

Contextual

"... lit spacecraft launching, set against a starry cosmic [-background.-] [+background at twilight.+]"

"In a futuristic setting, a lit city stands amidst a stark white landscape; above it swirls a galaxy of planets, moons, stars, and nebulas."

Content

"... above it swirls a galaxy of planets, moons, stars, and [+two+] nebulas."

Attribute

"In a futuristic setting, a lit city stands amidst a [-stark-] [+pristine+] white landscape;..."

Spatial

"... stark white landscape; [+around+] [-above-] it swirls a galaxy of planets, moons,..."

Contextual

"... stands amidst a stark white [-landscape-] [+landscape dotted with sparse vegetation+]; ..."

Figure 7: More Examples of prompt editing.

A.4 Prompt Modification Examples

In this subsection, we provide more examples of prompt modification such as the one in Figure 2, where a given prompt is modified based on our modification principles in Fig. 7.

A.5 Modification prompt instruction for LLM

To generate the negative prompts c^l for alignment training in our method, we have applied the Gemini 2.0 Flash model (Team et al., 2023) to modify the original prompts. Each modification is instructed to edit by using one or more editing strategies where each of these strategies following one of the modification principles described in Sec. 3.

We show how we instruct the Gemini 2.0 Flash model to modify the original prompt to generate $\boldsymbol{c}^l$. The left side of Fig. 8 shows how we instruct Gemini model to modify a given prompt. The right-hand side shows the choices of modification strategy, which correspond to the four modification principles.

Modification Instruction prompts

You are an assistant tasked with modifying the description prompt of a image so that the meaning of the two prompts are different, but the images generated from the two prompts are hard to distinguish.When provided with a text-to-image generation prompt, please modify the prompt and change only one element of the prompt using the following strategy.The goal here is to generate a modified prompt that is similar to the original prompt, but conceptually different, to trick the model to generate the same image
{strategy}

Please avoid color changes.
The image described by the modified text-to-image generation prompt and the original prompt should not be exactly the same, but visually similar.
Please output the modified text-to-image generation prompt, nothing else.
Here is the original text-to-image generation prompt:

{{prompt}}

Here is the modified text-to-image generation prompt:

Modification Strategies

Addition: Inserting new elements (objects, characters, landmarks) that weren't originally described.
Replacement/Substitution: Changing one element for another.
Quantitative Changes: Modifying the count or number of objects.
Appearance/Properties: Adjusting visual details such as texture, size, and material.
Style: Changing the artistic style or medium or adjusting descriptive adjectives to alter mood or tone.
Textual Detail: Enriching the description with additional adjectives or qualifiers to refine the prompt language and enhance realism or artistic expression.
Motion/Pose: Changing the dynamic state or pose of subjects.
Relations and Composition: Modifying spatial relationships among objects such as alignment, proximity, or composition.
Background: Altering or enriching the background elements.
Weather and Atmosphere: Introducing or modifying weather conditions.
Time and Lighting: Setting the scene at a specific time of day, season, or adjusting lighting conditions.
Setting: Changing the physical or cultural context

Figure 8: Left: the prompt template provide to the LLM to modify the a given prompt from image-prompt pair. Right: the precise way of modification that the LLM should apply for the given prompt.

A.6 t2i alignment's strong correlation with human preference

In Sec. 5, we demonstrated that the implicit preference score is strongly negatively correlated with all human preference score metrics. Intuitively, the implicit preference score is defined as the difference in diffusion loss between positive and negative image–prompt pairs. It quantifies how much more likely the model is to generate image $\boldsymbol{x}$ given its matching prompt $\boldsymbol{c}^w$ compared to the mismatched prompt $\boldsymbol{c}^l$. This loss, therefore, measures the model's ability to capture text–image alignment: models with higher implicit preference score also produce images that receive higher human preference scores. Consequently, **part of the human preference can be directly attributed to better text-image alignment**.

Figure 9: Left: Alice in a vibrant, dreamlike digital painting inside the Nemo Nautilus submarine. Right: An ethereal double portal with two paths illuminated by soft golden hour light, figures poised at the gateway in dynamic perspective, rendered in vibrant digital hues and sharp lines, with a serene mood and rich textures

Figure 10: Left: A glowing white fantasy castle with towers and spires, floating in a starlit sky. Right: Two androids hold hands, gazing into each other's eyes, in the ruins of a dystopian world, dramatic lighting.

To illustrate this, Figure 10 shows examples from the open-image-preferences-v1 dataset (Berenstein et al., 2024). Each pair compares two images generated from the same prompt: the left one contains the human-preferred image, and the right one the non-preferred image. Although all of these outputs are high-quality generations from state-of-the-art T2I models, annotators consistently favor the images that more faithfully reflect the prompt. For instance, in the first pair only the left image depicts the little girl "Alice." In the second pair the left image shows two portals, whereas the right shows only one. In the third pair only the left image appears to float in the sky. In the fourth pair the two androids hold hands in the left image but merely stand together in the right.

These qualitative examples, together with our quantitative ablation study of implicit preference score versus human preference scores, suggest that human judgments are closely tied to text–image alignment.

A.7 User study

We have conducted human study through user survey to evaluate the performance of our approach. An example of the user interface for our human study is shown in 11. During the study, we choose a subset of 100 prompts from PARTI-PROMPT dataset for evaluation. For each prompt, a pair of images generated by different methods from the prompt will be presented to the human subject. The human subject is then asked to select which image aligns better with the prompt, or if there is a tie. We gurantee that each pair of images are evaluated by at least 3 different human subjects.

A.8 Training dynamic

Training Curves of TKTO and TDPO As shown in Fig. 12, we plot the training dynamics of the Implicit Preference Score defined in Sec. 5, together with the curves of the alignment metrics Pick Score, CLIP, and Human Preference Score on a held-out validation set. Note that the TDPO loss objective in Eq. (10) can be written as $\mathcal{L}_{\text{TDPO}}(\theta) = -\mathbb{E}\big[\log\sigma\big(\alpha\,\text{IPS}(\theta) + c\big)\big]$, where $\alpha = \beta, \omega(t) > 0$ and c is a constant determined by

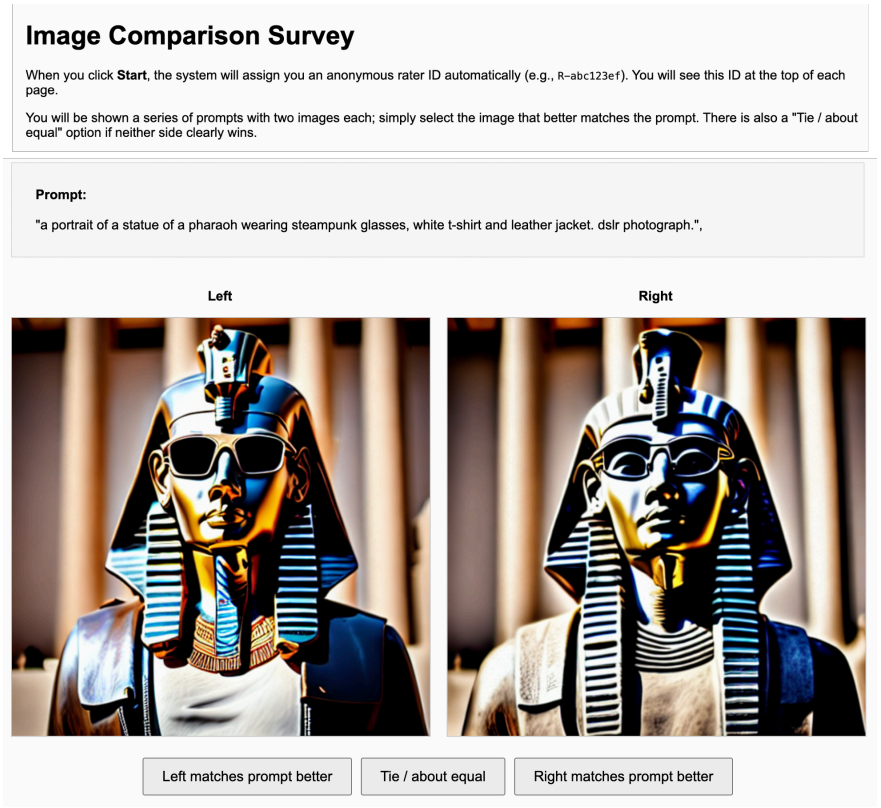

Figure 11: User study interface.

Figure 12: Training Curve of our TPO on SDXL. IPS stands for the Implicit Preferece Score.

the reference model, independent of θ. Since the log-sigmoid function $\log \sigma(\cdot)$ is monotonically increasing and $\alpha > 0$, minimizing the TDPO loss is equivalent to maximizing the Implicit Preference Score; consequently, as the training loss decreases, the Implicit Preference Score increases.

B More Experiment Results

B.1 clipping negative gradient

In our experiments, we observed that direct fine-tuning with objectives in the loss of TDPO and TKTO objectives would degrade the quality of the sample after several hundred steps. In our experiments, we found that directly optimizing towards the optimization objectives of TDPO and TKTO lead to unstable training. This is likely caused by the variance introduced by the negative gradient, a phenomenon extensively analyzed in prior work (Pal et al., 2024; Ren & Sutherland, 2025). To mitigate this, we introduce a clipping trick inspired by Proximal Policy Optimization (Schulman et al., 2017). Specifically, we clamp the squared L2

norm of the negative-sample term in the loss, which bounds extreme negative signals and stabilizes training. For example, in TDPO, we add a clamp function to the L2 squared norm on the θ parameter term condition on $\boldsymbol{c}^l$:

$$\begin{aligned} L_{\text{diff-tdpo}} = -\mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{c}^w,\boldsymbol{c}^l)\sim D, t\sim\mathcal{U}(0,T), \boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\log \sigma(\\ - \beta w(t)(||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{c}^w,t) - \boldsymbol{\epsilon}^w||_2^2 - ||\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t,\boldsymbol{c}^w,t) - \boldsymbol{\epsilon}^w||_2^2 \\ - (\text{clamp}(||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{c}^l,t) - \boldsymbol{\epsilon}^l||_2^2,\ \max = ||\boldsymbol{\epsilon}_{\mathbf{ref}}(\boldsymbol{x}_t,\boldsymbol{c}^l,t) - \boldsymbol{\epsilon}^l||_2^2 + \lambda_{\text{bound}}) \\ - ||\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t,\boldsymbol{c}^l,t) - \boldsymbol{\epsilon}^l||_2^2)))], \end{aligned} \tag{14}$$

where λ_{bound} is a hyper-parameter controls the strength of the maximum negative signal. For TKTO, this is applied similarly; the clamp function would be added when the condition prompt is not matched to the given image (when $\omega(\boldsymbol{c}) = -1$).

B.2 Human Preference Metric Performance

In addition to win rates, we also report the average scores of each model on each metric in a radar chart Fig. 13 for each dataset. We can see that our methods cover almost all the baselines on all metric dimensions and all dataset, suggesting that our methods have a higher average score for almost all the metric in all evaluation dataset.

B.3 More qualitative Comparison

In this subsection, we put more qualitative comparison images in Fig. 14

C The Use of Large Language Models (LLMs)

In this work, LLMs primarily serve as a prompt-editing tool. Details are provided in Appendix A.5. We also used them to improve writing by refining the flow of paragraphs and correcting grammar errors. It serves as an additional tool for polishing paper writing, but is not the main writer of this work.

D Limitation

Our preference-data-free alignment framework, while effective, has several limitations. First, its success hinges on the quality of prompt editing: our current strategies may overlook subtle semantic distinctions or produce unnatural negative examples. Second, because we rely on a budget-constrained off-the-shelf LLM, generated negatives can suffer from reduced fluency and faithfulness, which in turn can degrade alignment performance. Third, we fine-tune only the diffusion model while keeping the pre-trained text encoder fixed; this static encoder may limit the framework's capacity to discriminate between closely related prompts. Finally, generating all negatives from a single LLM restricts the diversity of hard negatives—incorporating adversarial, retrieval-based, or multi-model sampling strategies could further improve robustness.

E Reproducibility statement

Detail setup of our method, model architecture, training and evaluation pipeline has been outlined in the work. We also have a consistent training and evaluation setup for better reproduction. Moreover, the code for this work will be release and available soon.

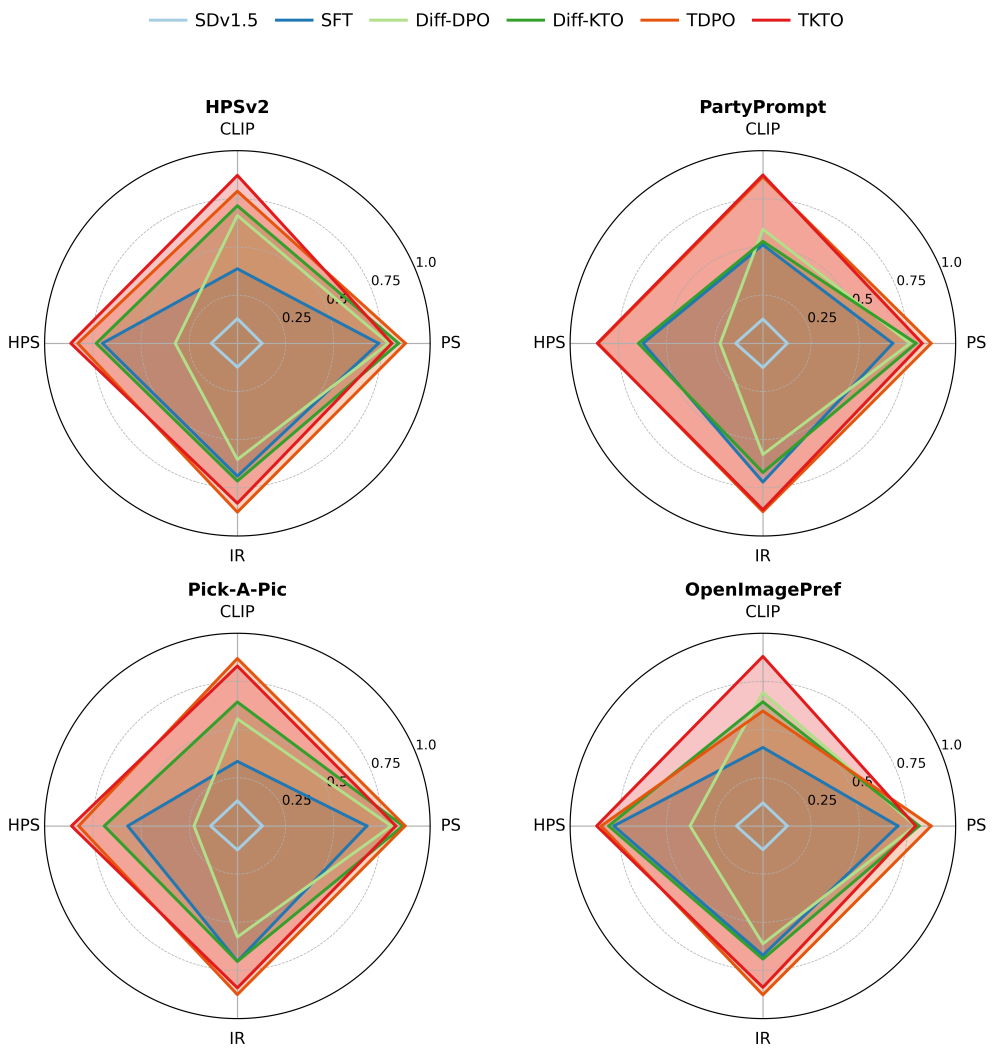

Figure 13: Radar chart of average scores of each model on each metric for our SD v1.5 experiments. The average scores are calculated by a relax min-max normalization and shifting the lower bound by one standard deviation to keep all curves in view. It shows that our methods (TDPO and TKTO) outperform the baselines (Diffusion-DPO, Diffusion-KTO, SFT and SD v1.5) on all metrics by covering all the baselines.

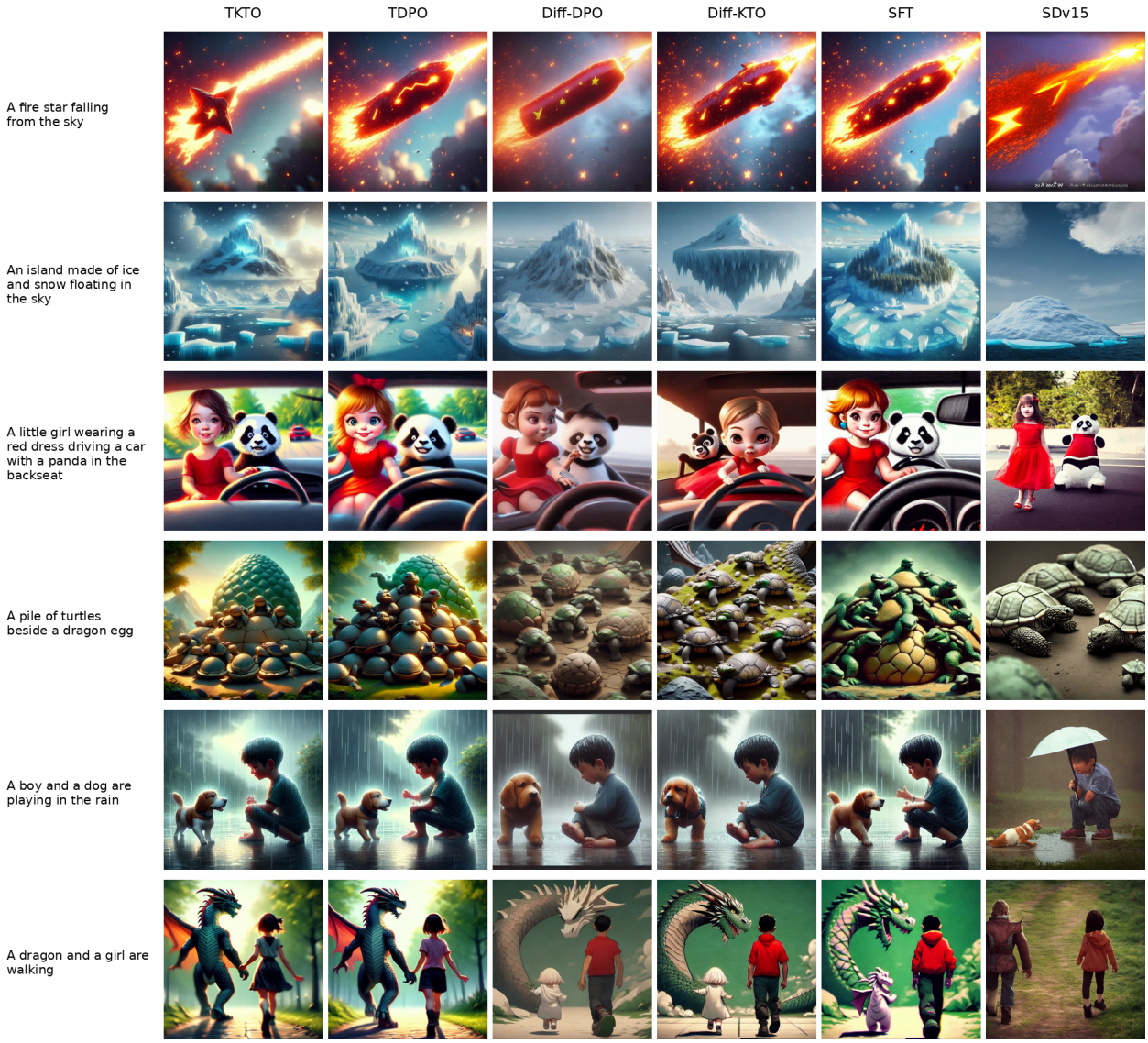

Figure 14: Side-by-Side grid comparison of the image generation using our methods and the baselines. The left most column show the prompts used to generate the images.