THE PREFERENCE IS IN THE DETAILS: TEXT-TO-IMAGE PREFERENCE ALIGNMENT WITH FINE-GRAINED VISUAL CUES

Pulkit Bansal
TCS Research, India
bansal.pulkit1@tcs.com

Vivek Srivastava
TCS Research, India
srivastava.vivek2@tcs.com

Shirish Karande
TCS Research, India
shirish.karande@tcs.com

ABSTRACT

Aligning text-to-image diffusion models with human preferences is essential for reliable deployment, yet existing approaches largely treat preference alignment as an output-level objective driven by coarse comparisons. Human visual judgment, however, is structured around fine-grained perceptual factors such as semantic coherence, texture fidelity, and local consistency, which require alignment at the level of internal representations. In this work, we present **PREFINE**, a framework that reformulates preference learning as a representational alignment problem by introducing structured, fine-grained preference supervision through controlled perturbations of high-quality images. These perturbations induce targeted variations along perceptually meaningful axes, encouraging diffusion models to develop representations that are sensitive to localized degradations while remaining robust to irrelevant variations. We further introduce a difficulty-aware curriculum that progressively refines perceptual sensitivity during training, enabling improved alignment with human judgments. Our experiments show that PREFINE consistently boosts alignment metrics across models and datasets, with gains in win rates of up to **13.0%** in Aesthetics Score and **15.2%** in ImageReward. These results suggest that fine-grained preference supervision improves alignment between learned visual representations and human perceptual evaluation, highlighting the role of structured preference signals in scalable alignment of generative models.

1 INTRODUCTION

Diffusion-based Text-to-Image (T2I) models have achieved impressive progress in generating high-quality images from textual prompts (Saharia et al., 2022; Betker et al., 2023; Nichol et al., 2021; Rombach et al., 2022b; Podell et al., 2023). These models are typically trained on large-scale datasets of text-image pairs and optimized using denoising objectives. However, despite their generative strength, such models (especially smaller, more lightweight variants like SD v1.5 that are commonly used in practical or resource-constrained settings) frequently fail to align with *fine-grained human preferences*. These preferences often involve subtle and localized attributes in semantics, structure, or aesthetics that significantly influence perceived image quality (Gu et al., 2024; Parihar et al., 2024; Liang et al., 2025; Wu et al., 2024a). Importantly, such judgments arise from sensitivity to fine-grained perceptual factors, suggesting that effective preference alignment requires not only improving outputs but also aligning the underlying visual representations with the perceptual structure of human evaluation.

Initial alignment efforts have largely relied on Reinforcement Learning from Human Feedback (RLHF) Black et al. (2023); Liang et al. (2024); Fan et al. (2023b), which trains reward models from curated human annotations and then fine-tunes the generation policy accordingly. However,

Figure 1: An example sequence of perturbations applied to the original image from the preference pair. The sequence is a combination of all 3 perturbation types: global, localized masked, and localized unmasked. We select the original and final perturbed image for preference expansion. The prompt for the original image is: "*Create a warm and inviting anime-style flyer to celebrate the 10-year anniversary of a mill. The event is aimed at individuals in the milling industry, including bakers, millers, and farmers. Feature cute, stylized characters in festive attire, a vibrant color palette, and a friendly, energetic atmosphere. Use flat vector rendering with sharp focus and dynamic composition.*"

RLHF pipelines are often computationally demanding, sensitive to reward model errors, and heavily reliant on high-quality human-labeled data (Casper et al., 2023; Rafailov et al., 2023). To reduce this dependency, reward-free approaches have emerged such as Diffusion-DPO Wallace et al. (2024), Diffusion-KTO Li et al. (2024b), and DSPO (Direct Score Preference Optimization) (Zhu et al.). These methods bypass reward modeling entirely by optimizing diffusion policies using pairwise comparisons. A more detailed background on these methods is provided in the Appendix A.

Yet both RLHF and DPO-style methods tend to treat preference alignment as a coarse global optimization problem by improving general image quality rather than learning how to distinguish and respond to *fine-grained visual cues*. In practice, however, overall image quality is not a singular concept. It emerges from the aggregation of multiple low- and mid-level visual factors: sharpness, compositional balance, semantic alignment, texture fidelity, and more. Capturing such nuanced preferences demands more than scaling up data; it demands designing *better supervision signals*. As a result, models may improve global quality metrics while remaining insensitive to localized perceptual deviations that strongly affect human judgment. This indicates that preference optimization alone does not necessarily induce representations that are aligned with the fine-grained perceptual axes underlying human preferences.

Building datasets that target these subtle factors is inherently challenging. Although recent works such as Sun et al., Hong et al. (2025), and Wu et al. (2024b) have explored synthetic data generation using pretrained models, such pipelines typically operate as black boxes with limited control. The resulting preferences often fail to exhibit interpretable or consistent preference signals, especially for fine-grained differences. As shown in Figure 6 in the Appendix, images generated by SD v1.5 and SDXL-Base under structured distortion prompts often fail to exhibit consistent degradation. This underscores the limited sensitivity of base models to subtle variations and the need for more controlled generation pipelines for reliable preference data. From a representational perspective, uncontrolled generation pipelines provide limited guarantees about which perceptual factors are being varied, making it difficult for models to learn consistent feature-level distinctions. This motivates the need for structured and controllable preference construction that can explicitly expose models to targeted perceptual variations.

In this work, we ask: Can fine-grained preference alignment be achieved by explicitly structuring supervision to better align diffusion model representations with human perceptual judgments, without relying on large-scale human annotation or uncontrolled black-box generations? Instead

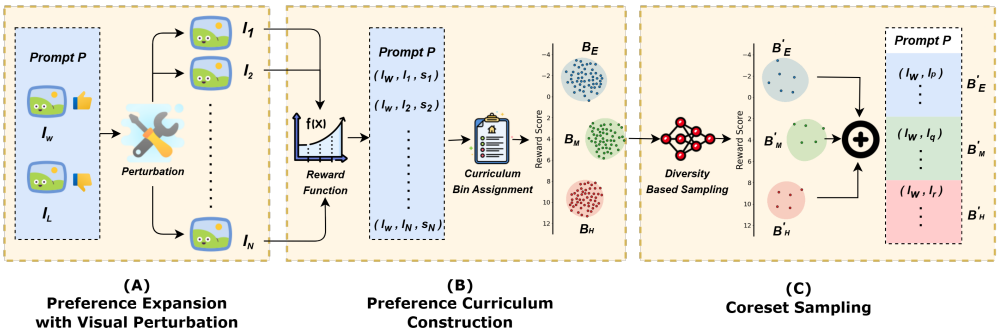


Figure 2: **Overview of the PREFINE Framework.** **(A)** Given a prompt P and a winning–losing image pair (I_w, I_l), we apply a suite of perturbations to generate synthetic losing candidates $(I_1, \ldots, I_N)$. **(B)** Each candidate is scored by a reward function $f(\cdot)$. The resulting triplets are sorted and binned into easy, medium, and hard levels to construct a difficulty-aware curriculum. **(C)** From each bin, the most diverse $M/3$ samples are selected using score dispersion and merged into a training batch of M preference-aligned pairs.

of scaling the size of preference data, we focus on scaling its *richness, structure, and controllability*. We introduce PREFINE, a framework for generating synthetic preferences that explicitly encode fine-grained visual distinctions (see Figure 2). Our method applies controlled perturbations to high-quality preference images by simulating realistic failure modes such as blur, compression artifacts, semantic drift, and compositional misalignment. These perturbed images are organized into a difficulty-based curriculum using a reward function, and sampled via a diversity-aware strategy to ensure broad coverage of visual nuances. These controlled perturbations induce systematic variations along perceptually meaningful dimensions, encouraging the model to develop representations that are sensitive to localized degradations while remaining robust to irrelevant variations.

By combining interpretability, curriculum structure, and targeted visual variation, PREFINE provides high-quality, fine-grained supervision that complements preference alignment techniques like Diffusion-DPO and DSPO. Inspired by recent work on perceptual artifact localization and multi-dimensional human preferences Zhang et al. (2023; 2024); Hendrycks et al. (2019), our perturbation strategy targets multiple aspects of image quality in a fine-grained manner that better reflects how humans evaluate generative outputs and promotes alignment between learned visual representations and human perceptual sensitivity. Additionally, the distortions we apply closely resemble failure patterns observed in real diffusion model generations Cao et al. (2024), making the resulting supervision more realistic and aligned with practical generative challenges.

Our key contributions are:

• **PREFINE Framework:** We propose PREFINE, a scalable framework that generates controllable, fine-grained synthetic preferences through structured perturbations that expose diffusion models to realistic failure modes, enabling fine-grained perceptual and representational alignment without expensive manual annotation or uncontrolled black-box generation.

• **Fine-grained Preference Curriculum:** We present a model-free *Preference Expansion* and *Curriculum Construction* technique to create a preference training curriculum, capturing fine-grained visual cues and progressively refining perceptual sensitivity during training.

• **Alignment Performance:** PREFINE improves alignment quality on standard diffusion models (e.g., SD v1.5, SDXL-Base) when combined with preference alignment methods like Diffusion-DPO and DSPO, making it model- and method-agnostic for text-to-image alignment.

2 RELATED WORK

2.1 PREFERENCE ALIGNMENT IN T2I GENERATION

Text-to-image (T2I) diffusion models have demonstrated impressive generative capabilities (Saharia et al., 2022; Betker et al., 2023; Ho et al., 2020; Nichol et al., 2021; Rombach et al., 2022b; Podell

et al., 2023). However, aligning their outputs with human preferences, particularly in subjective or open-domain prompts, remains a central challenge (Wallace et al., 2024; Xu et al., 2023). Reinforcement Learning from Human Feedback (RLHF) Christiano et al. (2017); Stiennon et al. (2020) has been a widely adopted approach and has recently been extended to diffusion models (Black et al., 2023; Liang et al., 2024; Fan et al., 2023b;a). Despite its success, RLHF introduces significant computational overhead and is susceptible to reward model inaccuracies and reward hacking.

To mitigate these limitations, several alternatives have been proposed that directly optimize preference objectives without explicit reward modeling. These include Diffusion-DPO Wallace et al. (2023), D3PO Yang et al. (2024), Diffusion-RPO Gu et al. (2024), Diffusion-KTO Li et al. (2024b), InPO Lu et al. (2025), DenseReward Yang et al., and DSPO (Zhu et al.). Similarly, SmPO-Diffusion Lu et al. models preference distributions to better account for variability in human judgments, while quality-aware pair ranking methods Singh et al. aim to improve alignment by selecting informative training pairs. PPD Dang et al. (2025) introduces a multi-reward optimization framework for personalized alignment using user-specific embeddings, and CaPO Lee et al. (2025) calibrates preferences across multiple reward models through frontier-based pair selection. These approaches improve scalability by relying on supervised learning with binary or pairwise preference data. However, they largely treat alignment as an output-level optimization problem, where improvements in preference scores do not necessarily imply improved sensitivity to fine-grained perceptual differences in the underlying representations. As a result, alignment quality remains strongly dependent on the scale, diversity, and structure of curated preference datasets.

2.2 Data Quality in Preference Supervision

Recent work has emphasized the critical role of dataset diversity and quality in effective preference alignment. Datasets such as *Pick-a-Pic v2* Kirstain et al. (2023), *ImageReward* Xu et al. (2023), and *HPD v2* Wu et al. (2023) rely on large-scale human-annotated pairwise comparisons. While effective, such datasets often exhibit limited coverage across styles, artifacts, or failure modes, which can introduce alignment bias and restrict the range of perceptual distinctions learned by models. To address these challenges, recent work has explored synthetic preference generation and automated filtering. For example, DreamSync Sun et al. and VisionPrefer Wu et al. (2024b) employ vision-language models or scoring functions to simulate preference feedback. Rich-HF Liang et al. (2024) focuses on collecting detailed misalignment annotations to improve reward modeling, while MaPO Hong et al. (2025) proposes controlled datasets targeting stylistic and safety-related preferences.

Although these approaches improve scalability, synthetic pipelines often operate as black boxes, providing limited control over which perceptual attributes vary across preference pairs. Consequently, the resulting supervision may not consistently expose models to structured variations along interpretable perceptual dimensions. In contrast, our approach emphasizes *fine-grained preference supervision* through controlled perturbations applied to high-quality images. By systematically introducing localized and interpretable degradations, our framework constructs preference signals that explicitly vary along perceptually meaningful axes. This structured supervision enables models to learn consistent distinctions across subtle visual factors, encouraging improved alignment between learned visual representations and human perceptual judgments while maintaining minimal annotation overhead.

2.3 Curriculum Learning for Image Generation

Curriculum learning was introduced by Bengio et al. (2009), who demonstrated that training models on examples ordered from easy to hard can accelerate convergence and improve generalization. This idea has since been extended to image generation and generative modeling. Curriculum GANs Sharma et al. (2018) improve stability by gradually increasing task difficulty, such as progressively increasing image resolution. Image Difficulty Curriculum GAN (CuGAN) Soviany et al. (2020) ranks images using learned difficulty scores to structure training through progressively harder samples and adaptive sampling strategies. More recent work has explored curriculum-based optimization in diffusion models and alignment settings. Curriculum DPO Croitoru et al. (2025) ranks generated samples using reward functions and selects pairs with progressively smaller preference gaps, while Curry-DPO Pattnaik et al. (2024) introduces curriculum-based ordering of preference

pairs for language model alignment. ObjBlur Frolov et al. (2024) proposes a curriculum for layout-to-image generation by progressively reducing layout ambiguity during training.

In contrast to prior approaches that infer difficulty through learned reward signals or ranking heuristics, PREFINE explicitly engineers its curriculum through structured perturbations. Difficulty is determined by controlled variations in perceptual degradation rather than naive post hoc scoring, enabling progressive exposure to fine-grained visual differences. This design encourages models to gradually refine perceptual sensitivity, supporting alignment not only at the output level but also at the level of representations underlying human visual evaluation.

3 OUR APPROACH

In this section, we discuss our proposed framework PREFINE in detail, which promotes T2I preference alignment to focus on capturing fine-grained visual details (see Appendix B for a summary of the notation used).

3.1 PROBLEM FORMULATION

We consider a dataset $\mathcal{T} = \{(\mathcal{P}, I_w, I_l)\}$, where each sample consists of a text prompt $\mathcal{P}$, a winning image I_w, and a losing image I_l. To support learning of fine-grained visual preferences, we aim to construct an expanded dataset $\mathcal{T}'$ with perturbed image pairs that reflect subtle degradations. We define $\mathcal{D}$, a set of perturbation functions, to generate perturbed candidates $\{I_1, I_2, \ldots, I_N\}$, and construct the expanded set $\mathcal{T}'$ accordingly. This enables generating difficulty-aware preference pairs for training models sensitive to subtle quality differences, as defined in Equation 1.

$$\mathcal{T}' = \{(\mathcal{P}, I_w, I_j) \mid j \in \{1, 2, \ldots, N\},\ I_w \succ I_j\} \tag{1}$$

Next, we present our framework PREFINE and discuss the three main components in detail. We present an overview of our framework in Algorithm 1 and Figure 2.

3.2 PREFERENCE EXPANSION WITH VISUAL PERTURBATION

We first expand the preference sample by synthesizing losing candidates through the injection of fine-grained visual cues into the original preference pair. Specifically, given a prompt $\mathcal{P}$, a winning image I_w, and a losing image I_l ($I_w, I_l \in \mathbb{R}^{H\times W\times 3}$), we generate a total of N perturbed candidates from both images combined. The prompt $\mathcal{P}$ and the winning image I_w are preserved across all augmentations, while each perturbed candidate $\{I_1, I_2, \ldots, I_N\}$ becomes a new losing image (see Algorithm 2).

Each variant I_i is generated by sequentially composing multiple perturbation operations. Instead of applying a single transformation, we sample and chain perturbations from different categories, progressively modifying the image (see Figure 1 and Appendix). This process produces diverse distortions with varied visual effects. Perturbations are drawn from three spatial categories: *localized masked* (e.g., swirl or twist within a soft mask), *localized unmasked* (e.g., pixelation on a patch), and *global* (e.g., Gaussian noise or compression over the full image).

1. Localized masked perturbations. These operations are applied within spatially constrained regions defined by a binary mask $M \in \{0, 1\}^{H\times W}$, computed via the convex hull of randomly sampled control points (ope; Lee, 1983). To enable smooth transitions, the binary mask is blurred using a Gaussian kernel to produce a soft alpha mask $\alpha \in [0, 1]^{H\times W}$ as:

$$\alpha = \text{GaussianBlur}(M, \sigma) \tag{2}$$

where `GaussianBlur` denotes a spatial convolution with a Gaussian kernel. This mask softly blends between the perturbed and original image. Given a perturbation function $D(\cdot)$, the final blended output is:

$$\tilde{I} = \alpha \odot D(I) + (1 - \alpha) \odot I \tag{3}$$

where $\odot$ denotes element-wise multiplication. Perturbations in this category include *swirl*, *twist*, *radial zoom*, and *sine wave*, which introduce localized geometric changes while preserving global structure.

Algorithm 1 PREFINE

Require: Dataset $\mathcal{T} = \{(\mathcal{P}, I_w, I_l)\}$, distortions $\mathcal{D}$, scorer $f(\cdot)$, variants N, target size M
Ensure: Curriculum-ordered dataset $\mathcal{D}_{\text{train}}$

$\mathcal{C} \leftarrow \emptyset$ {Candidate triplets}
Let the input be $X = (\mathcal{P}, I_w, I_l)$
for all $X \in \mathcal{T}$ **do**
  for $i = 1$ to N **do**
    $\tilde{I}_i \leftarrow$ PREFERENCEEXPANSION$(X, \mathcal{D})$
    $s_i \leftarrow f(\tilde{I}_i)$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{(I_w, \tilde{I}_i, s_i)\}$
  end for
end for
$\mathcal{B}_{ALL} \leftarrow$ PREFERENCECURRICULUM$(\mathcal{C})$
for all $\mathcal{B} \in \mathcal{B}_{ALL}$ **do**
  $\mathcal{B}' \leftarrow \mathcal{B} \cup$ CORESETSAMPLING$(\mathcal{B}, M/3)$
end for
return $\mathcal{B}'$

Algorithm 2 PREFERENCE EXPANSION

Require: Image I, distortion set $\mathcal{D}$
Ensure: Distorted image $\tilde{I}$

Randomly select distortion $D \in \mathcal{D}$
if D is **masked then**
  $M \leftarrow$ CONVEXHULL(RandomPoints)
  $\alpha \leftarrow$ GAUSSIANBLUR(M)
  $\tilde{I} = \alpha \odot D(I) + (1 - \alpha) \odot I$
else if D is **unmasked then**
  Sample $R = [x_1 : x_2, y_1 : y_2]$
  Extract patch $P = I[R]$
  Apply distortion to patch: $\tilde{P} = D(P)$
  Replace region: $I[R] \leftarrow \tilde{P}$
  $\tilde{I} \leftarrow I$
else
  $\tilde{I} = D(I)$ {Apply to entire image}
end if
return $\tilde{I}$

Algorithm 3 PREFERENCE CURRICULUM

Require: Candidate set $\mathcal{C} = \{(I_w, \tilde{I}, s)\}$, number of bins $k = 3$
Ensure: Binned subsets $\{\mathcal{B}_E, \mathcal{B}_M, \mathcal{B}_H\}$

$n \leftarrow |\mathcal{C}|$ {Get total candidates}
Sort $\mathcal{C}$ by difficulty score s_i (in the ascending order)
{Determine bin boundaries}
$n_1 \leftarrow \lfloor n/3 \rfloor$ {Easy bin cutoff}
$n_2 \leftarrow \lfloor 2n/3 \rfloor$ {Medium bin cutoff}
$\mathcal{B}_E \leftarrow \mathcal{C}[0 : n_1]$ {Create bin for Easy complexity samples}
$\mathcal{B}_M \leftarrow \mathcal{C}[n_1 : n_2]$ {Create bin for Medium complexity samples}
$\mathcal{B}_H \leftarrow \mathcal{C}[n_2 : n]$ {Create bin for Hard complexity samples}
$\mathcal{B}_{ALL} = \{\mathcal{B}_E, \mathcal{B}_M, \mathcal{B}_H\}$
return $\mathcal{B}_{ALL}$

Algorithm 4 CORESET SAMPLING

Require: Triplet bin $\mathcal{B} = \{(I_w, \tilde{I}_i, s_i)\}_{i=1}^{N}$, target size K
Ensure: Diverse subset $\mathcal{B}' \subset \mathcal{B}$ of size K

Sort $\mathcal{B}$ by s_i; $V \leftarrow$ sorted triplets, $S \leftarrow$ scores
Init 3D cache $\texttt{dp}[p][r][l]$ for $0 \le p < N$, $0 \le r \le K$, $-1 \le l < N$
Define recursive DP(p, r, l):
if $r = 0$ **then**
  return 0
else if $p = N$ **then**
  return $-\infty$
else
  $\delta \leftarrow |S_p - S_l|$ if $l \neq -1$, else 0
  $a \leftarrow$ DP$(p + 1, r, l)$ {skip}
  $b \leftarrow \delta +$ DP$(p + 1, r - 1, p)$ {take}
  return $\max(a, b)$
end if
Run DP$(0, K, -1)$ and backtrack to get $\mathcal{B}'$
return $\mathcal{B}'$

2. Localized unmasked perturbations. These perturbations operate on a randomly sampled rectangular region $R = [x_1 : x_2,\, y_1 : y_2] \subset [1, H] \times [1, W]$, without using any masks or blending. A sub-region P is extracted as:

$$P = I[x_1 : x_2,\, y_1 : y_2] \tag{4}$$

and transformed directly:

$$\tilde{P} = D(P), \text{ followed by } I[x_1 : x_2,\, y_1 : y_2] \leftarrow \tilde{P} \tag{5}$$

These operations simulate occlusions, noise, or localized disruptions. Functions of this type include *pixelation*, *erase with inpainting*, and *color jitter*.

3. Global perturbations. These transformations are applied over the entire image without any spatial constraint. The operation is defined simply as:

$$\tilde{I} = D(I), \tag{6}$$

where D is a global transformation function. This category includes perturbations like *Gaussian blur*, *Gaussian noise*, *salt-and-pepper noise*, *channel swapping*, *shearing*, *posterization*, *JPEG compression*, and *elastic transformation*.

We present the example of the perturbation functions used in Figures 1. By combining the three categories: soft masked localized, region based unmasked, and global perturbations, we generate a rich set of challenging candidates. The visual diversity supports learning of spatially aware and perceptually robust models sensitive to fine-grained visual differences. See the Appendix C.1 for a detailed description of the perturbation types.

3.3 Preference Curriculum Construction

Preference expansion provides us with prompt $\mathcal{P}$, one winning image I_w and N losing images $I_1, I_2, \ldots, I_N$. To enable quantitative comparison, we project each image onto a numerical scale using a reward function $f : I \to \mathbb{R}$, where f maps an input image to a real-valued score. Applying f to each losing image yields scores $s_1, s_2, \ldots, s_N$, corresponding to the images $I_1, I_2, \ldots, I_N$, respectively. Using the winning image I_w, the perturbed losing images $I_1, I_2, \ldots, I_N$, and their associated scores $s_1, s_2, \ldots, s_N$, we construct N triplets of the form $(I_w, I_1, s_1), (I_w, I_2, s_2), \ldots, (I_w, I_N, s_N)$. We denote the resulting set of all triplets as:

$$\mathcal{B} = \{(I_w, I_i, s_i)\}_{i=1}^{N} \tag{7}$$

These triplets are sorted in ascending order based on the score s_i of the losing image and divided into r bins as $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \ldots, \mathcal{B}_r$ with different sizes $k_1, k_2, k_3, \ldots, k_r$. For our experiments, we use $r = 3$ and all bins have equal size (i.e., $k_1 = k_2 = k_3 = \ldots = k_r$), and in this way we obtain three bins: $\mathcal{B}_E$, $\mathcal{B}_M$, and $\mathcal{B}_H$. The detailed procedure is described in Algorithm 3.

- **Easy complexity bin** (B_E): Triplets where the losing images have the lowest scores.
- **Medium complexity bin** (B_M): Triplets with mid-range losing scores, representing moderate difficulty.
- **Hard complexity bin** (B_H): Triplets where the losing images have the highest scores, making them harder to distinguish from the winning image.

3.4 Coreset Sampling

After constructing the curriculum, our goal is to select the top M diverse triplets from the available N, following the principle of *coreset sampling*, which involves choosing a small yet informative subset that preserves key data characteristics. In our case, *diversity* is the measure of informativeness. To ensure this, we adopt a score-based strategy: from each of three bins, we select $M/3$ triplets using the *Coreset Sampling* (see Algorithm 4), which maximizes the overall score dispersion in the selected subset.

This selection problem can be viewed as a variant of the classic 0/1 Knapsack problem Laabadi et al. (2018); Martello & Toth (1990); Gawiejnowicz et al. (2023). Each item (triplet) has unit weight, and we must select exactly $M/3$ items, similar to a fixed-capacity knapsack. However, unlike the standard setting where item values are fixed, here the *value* is dynamic and depends on the the absolute difference from the current maximum score among selected items. This interaction-based, state-dependent value structure makes it a natural extension of the knapsack problem. See the Appendix D.1 for a detailed analogy.

We construct the final set $\mathcal{B}' \subset \mathcal{B}$ of size M by selecting $M/3$ triplets from each bin. For each $\mathcal{B}_k \in \{\mathcal{B}_E, \mathcal{B}_M, \mathcal{B}_H\}$, we form an ordered subset $\mathcal{B}'_k \subset \mathcal{B}_k$ such that:

$$\mathcal{B}'_k = \underset{|\mathcal{B}'_k|=M/3}{\arg\max} \sum_{j=1}^{M/3-1} |s_{j+1} - s_j| \tag{8}$$

Finally, we form training pairs using the selected triplets by pairing each winning image I_w with its corresponding losing image. The selected samples from the Easy, Medium, and Hard bins are merged in order of increasing difficulty, beginning with $\mathcal{B}_E$, then $\mathcal{B}_M$, and finally $\mathcal{B}_H$. The final training subset is then constructed as:

$$\mathcal{B}' = \mathcal{B}'_E \cup \mathcal{B}'_M \cup \mathcal{B}'_H, \quad \text{with } |\mathcal{B}'| = M \tag{9}$$

This results in a total of M image pairs, which are used to create a training batch of size M. Each pair maintains the preference structure and contributes to a curriculum-aligned training process.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset and Models: We use the `open-image-preferences-v1-binarized` hug (b) dataset, comprising 7,459 human-annotated image preference pairs. In this dataset, all images are generated using two models: `FLUX.1-dev` and `stable-diffusion-3.5-large` (hug, a; Esser et al., 2024). We apply our PREFINE method to this dataset to generate a new, enriched dataset. We use *ImageReward* as a reward model during data generation. Similar to Diffusion-DPO and DSPO, for training and evaluation, we employ two widely used diffusion models: *SD v1.5* and *SDXL-Base* (Rombach et al., 2022a; Podell et al., 2023).

Baselines: PREFINE aims to generate a diverse curriculum-based dataset for preference alignment. To evaluate its effectiveness, we train existing alignment techniques, including Supervised Fine-Tuning (SFT), Diffusion-DPO, and DSPO, using both the original human preference dataset and our PREFINE-generated dataset. We also include the pretrained SD v1.5 and SDXL-Base as baselines.

Evaluation: To assess human preference alignment, we perform T2I generation. We use test prompts from Pick-a-pic V2 Kirstain et al. (2023), GenAI-Bench Li et al. (2024a), and PartiPrompts Yu et al. (2022). For quantitative analysis, we use Aesthetics Score Schuhmann (2022) and ImageReward Xu et al. (2023).

Implementation Details: For each human-annotated image pair, we generate $N{=}150$ perturbed images using 3–11 perturbations randomly selected from a set of 16 perturbations, and select $M = \{4, 8, 16\}$ diverse images. Each sample is scored using *ImageReward*. The choice of M corresponds to the batch sizes used during fine-tuning. We apply LoRA-based parameter-efficient fine-tuning with batch sizes of 4, 8, and 16. Training is executed for 350 checkpoints in the case of Diffusion-DPO, and 1000 checkpoints for both DSPO and SFT. Hyperparameter details are provided in the Appendix E.1.

5 RESULTS

5.1 QUANTITATIVE RESULTS

Our method consistently improves both SD v1.5 and SDXL-Base across subjective and reward-based metrics, measured using win rates, which represent the percentage of samples where one method outperforms the base model. On SD v1.5 (Table 1a), we observe notable gains in Aesthetics score and ImageReward win rates. For example, on Pick-a-Pic V2, Ours+Diffusion-DPO surpasses Diffusion-DPO by over **6%** in both metrics. On GenAIBench, Ours+DSPO achieves win rate improvements of **+5.2%** for Aesthetics score and **+4.1%** for ImageReward. With SDXL-Base (Table 1b), On Pick-a-Pic V2, Ours+Diffusion-DPO achieves an ImageReward win rate of **64.8%**, representing a **+15.2%** improvement over Diffusion-DPO. Similarly, on GenAIBench, Ours+Diffusion-DPO delivers win rate improvements of **+6.6%** in Aesthetics score and **+11.0%** in ImageReward. These results shows that our method complements existing preference optimization techniques like Diffusion-DPO and DSPO, providing significant improvements across models and benchmarks.

5.2 QUALITATIVE ANALYSIS

We present the results for SD v1.5 using Diffusion-DPO in Figure 4 . For the prompt *"A cute wooden owl statue holding a large globe of the Earth above its head"*, The base model captures a rough shape, and Diffusion-DPO improves form but lacks clarity; in contrast, Ours+Diffusion-DPO generates a crisply rendered owl with clear wooden textures and a correctly placed globe. Similarly, for *"Two cats playing with a single ball"*, the base model renders two balls, and Diffusion-DPO fails to capture natural interaction, while Ours+Diffusion-DPO accurately depicts both cats and a single shared object with correct relational context. These examples show how preference tuning with PREFINE enhances visual grounding, object relationships, and compositional quality. We further

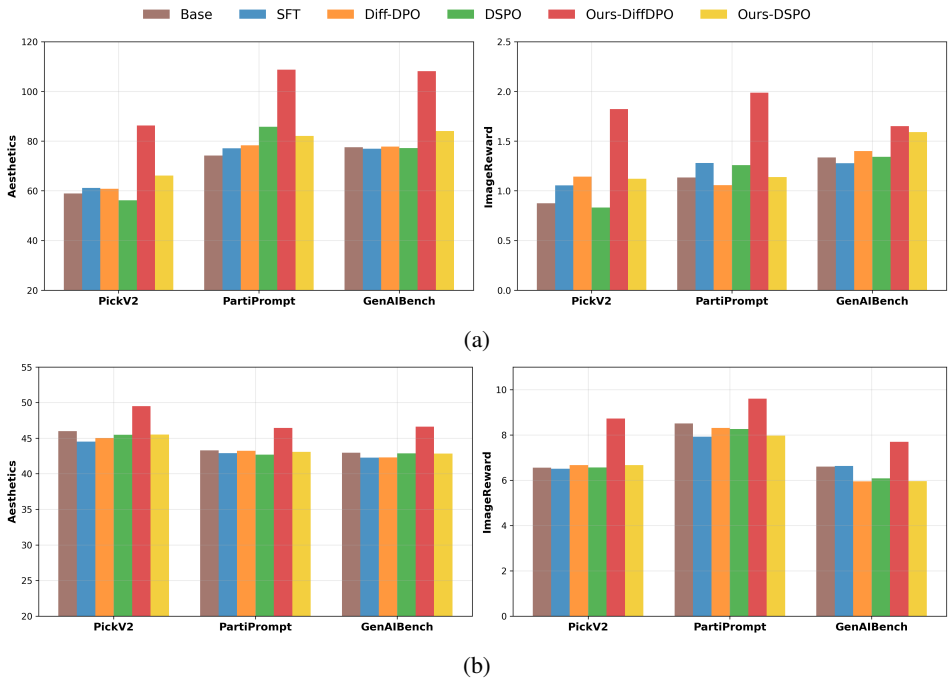
Table 1: Win rate percentage against the base models SD v1.5 (Table 1a) and SDXL-Base (Table 1b). Best results are highlighted in **boldface**. Second best results are underlined. Values in parentheses show performance change with PREFINE for each alignment method. Rows for our method are shaded.

Dataset	Method	Metrics	
		Aesthetics	ImageReward
Pick-a-Pic V2	SFT	51.4	54.2
	Diff.-DPO	48.4	54.6
	DSPO	46.2	48.8
	Ours + Diff.-DPO	**58.6** (+10.2 ↑)	**63.0** (+8.4 ↑)
	Ours + DSPO	49.6 (+3.4 ↑)	53.0 (+4.2 ↑)
PartiPrompt	SFT	52.2	49.2
	Diff.-DPO	53.0	47.0
	DSPO	56.4	49.8
	Ours + Diff.-DPO	**65.8** (+12.8 ↑)	**58.6** (+11.6 ↑)
	Ours + DSPO	51.8 (-4.6 ↓)	52.6 (+2.8 ↑)
GenAIBench	SFT	52.6	47.8
	Diff.-DPO	49.6	49.2
	DSPO	48.4	49.6
	Ours + Diff.-DPO	**62.6** (+13.0 ↑)	51.8 (+2.6 ↑)
	Ours + DSPO	52.6 (+4.2 ↑)	**52.8** (+3.2 ↑)

(a) SD v1.5

Dataset	Method	Metrics	
		Aesthetics	ImageReward
Pick-a-Pic V2	SFT	48.0	46.6
	Diff.-DPO	49.0	49.6
	DSPO	48.2	49.2
	Ours + Diff.-DPO	**54.2** (+5.2 ↑)	**64.8** (+15.2 ↑)
	Ours + DSPO	49.4 (+1.2 ↑)	51.2 (+2.0 ↑)
PartiPrompt	SFT	47.8	47.2
	Diff.-DPO	50.6	47.6
	DSPO	48.4	47.4
	Ours + Diff.-DPO	**55.6** (+5.0 ↑)	**58.0** (+10.4 ↑)
	Ours + DSPO	49.0 (+0.6 ↑)	49.8 (+2.4 ↑)
GenAIBench	SFT	46.2	51.2
	Diff.-DPO	51.2	48.4
	DSPO	49.6	47.6
	Ours + Diff.-DPO	**57.8** (+6.6 ↑)	**59.4** (+11.0 ↑)
	Ours + DSPO	50.4 (+0.8 ↑)	49.0 (+1.4 ↑)

(b) SDXL-Base



(a)

(b)

Figure 3: Quantitative comparison of different fine-tuning methods on **(a) SD v1.5** and **(b) SDXL-Base** across three datasets-Pick-a-Pic V2, PartiPrompt, and GenAIBench, using three evaluation metrics: Aesthetics score(center), and ImageReward (right). All metrics are rescaled for better visualization.

evaluate PREFINE's text rendering capability by adding 200 samples from the MARIO-Eval Chen et al. (2023) dataset into the training set. As shown in Figure 5, PREFINE improves text legibility over the base model and Diffusion-DPO.

6 CONCLUSION

In this work, we addressed the challenge of aligning text-to-image diffusion models with fine-grained human preferences, which is often limited by the structure and diversity of existing preference data. We introduced PREFINE, a scalable framework that enriches preference supervision

Figure 4: Qualitative performance of preference alignment of SD v1.5 with PREFINE using Diffusion-DPO. We present two prompts from each dataset. The fine-grained details in the generated images and the aesthetic quality have improved significantly with PREFINE.

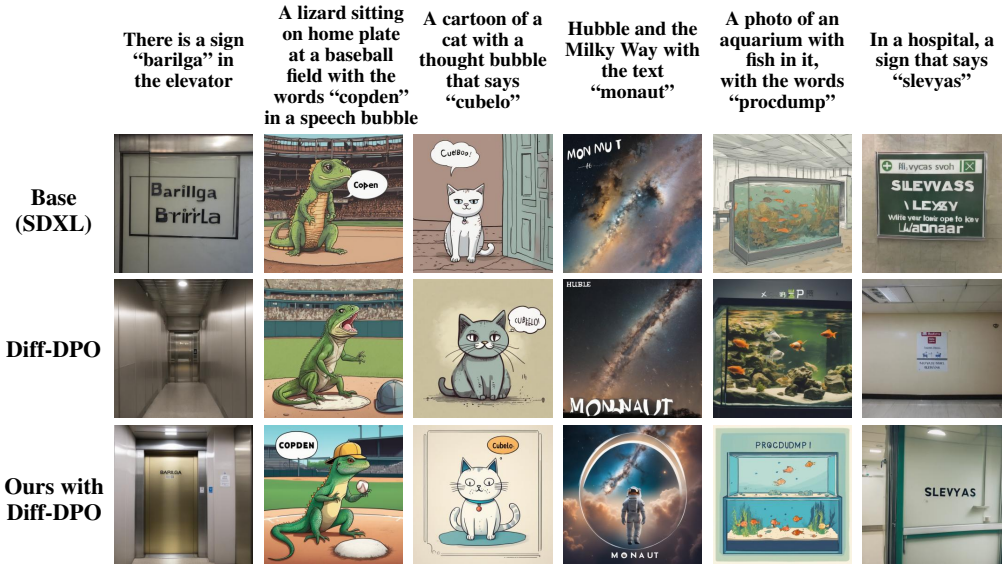


Figure 5: Qualitative comparison of text rendering in SDXL-Base with PREFINE using Diffusion-DPO. Prompts are from CreativeBench Yang et al. (2023).

through controlled perturbations, enabling the construction of fine-grained and interpretable preference signals without large-scale human annotation. By exposing models to structured variations along perceptually meaningful dimensions, PREFINE improves sensitivity to localized visual degradations, leading to better compositional understanding, spatial consistency, and fine-grained detail rendering. Our results highlight that effective preference alignment depends not only on the optimization objective but also on how supervision shapes perceptual distinctions learned by the model. By adopting a data-centric approach, PREFINE complements existing preference optimization methods and provides a practical pathway toward scalable fine-grained alignment in generative models.

REFERENCES

black-forest-labs/FLUX.1-dev · Hugging Face — huggingface.co. https://huggingface.co/black-forest-labs/FLUX.1-dev, a. [Accessed 10-05-2025].

data-is-better-together/open-image-preferences-v1-binarized · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/data-is-better-together/open-image-preferences-v1-binarized, b. [Accessed 10-05-2025].

OpenCV: OpenCV modules — docs.opencv.org. https://docs.opencv.org/4.x/index.html. [Accessed 18-07-2025].

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

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 (2023). *URL https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Bin Cao, Jianhao Yuan, Yexin Liu, Jian Li, Shuyang Sun, Jing Liu, and Bo Zhao. Synartifact: Classifying and alleviating artifacts in synthetic images via vision-language model. *arXiv preprint arXiv:2402.18068*, 2024.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2824–2834, 2025.

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8020–8030, 2025.

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*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023a.

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, 2023b.

Stanislav Frolov, Brian Moser, Sebastian Palacio, and Andreas Dengel. Objblur: A curriculum learning approach with progressive object-level blurring for improved layout-to-image generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10621–10629, 2024.

Stanisław Gawiejnowicz, Nir Halman, and Hans Kellerer. Knapsack problems with position-dependent item weights or profits. *Annals of Operations Research*, 326(1):137–156, 2023.

Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Soukaina Laabadi, Mohamed Naimi, Hassan El Amri, Boujemâa Achchab, et al. The 0/1 multi-dimensional knapsack problem and its variants: A survey of practical models and heuristic approaches. *American Journal of Operations Research*, 8(05):395, 2018.

Der-Tsai Lee. On finding the convex hull of a simple polygon. *International journal of computer & information sciences*, 12(2):87–98, 1983.

Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18465–18475, 2025.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024a.

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

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13199–13208, 2025.

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. In *Forty-second International Conference on Machine Learning*.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. Inpo: Inversion preference optimization with reparametrized ddim for efficient diffusion model alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28629–28639, 2025.

Silvano Martello and Paolo Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*, pp. 469–487. Springer, 2024.

Pulkit Pattnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. Curry-dpo: Enhancing alignment using curriculum learning & ranked preferences. *arXiv preprint arXiv:2403.07230*, 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. *arXiv preprint arXiv:2307.01952*, 2023.

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.

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 (CVPR)*, pp. 10684–10695, June 2022a.

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, 2022b.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann. LAION-Aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022. Accessed: 2023-11-10.

Rishi Sharma, Shane Barratt, Stefano Ermon, and Vijay Pande. Improved training with curriculum gans. *arXiv preprint arXiv:1807.09295*, 2018.

Kunal Singh, Mukund Khanna, and Pradeep Moturi. Effective text-to-image alignment with quality aware pair ranking. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu. Image difficulty curriculum for generative adversarial networks (cugan). In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3463–3472, 2020.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 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.

Tong Wu, Yinghao Xu, Ryan Po, Mengchen Zhang, Guandao Yang, Jiaqi Wang, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Fiva: Fine-grained visual attribute dataset for text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:31990–32011, 2024a.

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. *CoRR*, 2023.

Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Multimodal large language models make text-to-image generative models align better. *Advances in Neural Information Processing Systems*, 37:81287–81323, 2024b.

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, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Forty-first International Conference on Machine Learning*.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7579–7590, 2023.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8018–8027, 2024.

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*.

A BACKGROUND

A.1 DENOISING DIFFUSION PROBABILISTIC MODELS

Denoising Diffusion Probabilistic Models Ho et al. (2020) define a forward noising process that incrementally corrupts a clean image $\mathbf{x}_0 \sim p_{\text{data}}$ in T discrete steps via a fixed variance schedule $\{\beta_t\}_{t=1}^{T}$. In each step:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\, \beta_t \mathbf{I}\big), \tag{10}$$

so that overall

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}). \tag{11}$$

The model then learns the reverse Markov chain $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, parameterized by a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ that predicts the added noise:

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) &= \mathcal{N}\Big(\mathbf{x}_{t-1}; \frac{\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1-\beta_t}}, \sigma_t^2 \mathbf{I}\Big), \\ \text{where } \sigma_t^2 &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \qquad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s). \end{aligned} \tag{12}$$

Training is performed by minimizing the re-weighted mean-squared error

$$L(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon}\Big[\lambda(t)\, \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\Big], \tag{13}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

A.2 RLHF FOR TEXT-TO-IMAGE DIFFUSION MODELS

Reinforcement Learning from Human Feedback (RLHF) adapts a pretrained diffusion model using human preferences. First, a reward model $r(c, x)$ is trained on comparison data (x_w, x_l) using a Bradley–Terry (BT) likelihood:

$$P(x_w \succ x_l \mid c) = \sigma\big(r(c, x_w) - r(c, x_l)\big), \tag{14}$$

where c is the text prompt and σ the sigmoid function. Training minimizes the negative log-likelihood of observed preferences.

Next, the diffusion model, viewed as a multi-step Markov Decision Process (MDP) producing trajectories $(x_{0:T})$, is fine-tuned to maximize cumulative reward while staying close to its pretrained behavior:

$$\mathcal{L}_{\text{RLHF}} = \mathbb{E}_{c,\, x_{0:T} \sim p_\theta(\cdot|c)}\left[\sum_{t=0}^{T-1} r(x_t, c)\right] - \lambda\, D_{\text{KL}}\big(p_\theta(x_{0:T} \mid c) \,\|\, p_{\text{pref}}(x_{0:T} \mid c)\big) \tag{15}$$

where p_{pref} is the pretrained diffusion distribution, and λ balances reward versus KL regularization.

A.3 Preference Alignment with Diffusion DPO

Diffusion-DPO Wallace et al. (2023) adapts Direct Preference Optimization to diffusion models by defining a reward on the full reverse trajectory and optimizing for alignment with human preferences, without requiring a reward model.

Trajectory-level reward. Given a prompt c and a preferred sample $x_0^w \succ x_0^l$, we define the preference-based reward on the clean image x_0 as the expected reward over the reverse diffusion trajectory $x_{1:T} \sim p_\theta(\cdot \mid x_0, c)$:

$$r(c, x_0) = \mathbb{E}_{x_{1:T} \sim p_\theta(\cdot|x_0,c)} \left[R(c, x_{0:T}) \right], \tag{16}$$

where $R(c, x_{0:T})$ is a reward function over the entire denoising path.

Final Diffusion-DPO training objective. Using a preference dataset $D = \{(c, x_0^w, x_0^l)\}$, the loss encourages higher reward for preferred samples over less preferred ones. After approximation and simplification, the final training loss becomes:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x_0^w, x_0^l) \sim D,\ t \sim \mathcal{U}(0,T)} \log \sigma \Big(-\beta T\, \omega(\lambda_t) \Big[\|\epsilon^w - \epsilon_\theta(x_t^w, t)\|^2 - \|\epsilon^w - \epsilon_{\text{ref}}(x_t^w, t)\|^2 - \|\epsilon^l - \epsilon_\theta(x_t^l, t)\|^2 + \|\epsilon^l - \epsilon_{\text{ref}}(x_t^l, t)\|^2 \Big] \Big)$$

where $x_t^* = \alpha_t x_0^* + \sigma_t \epsilon^*$, with $\epsilon^* \sim \mathcal{N}(0, I)$, and $\lambda_t = \alpha_t^2 / \sigma_t^2$ is the signal-to-noise ratio at step t. The loss aligns the model's denoising trajectory with human preferences, relative to a frozen reference model.

A.4 Preference Alignment with DSPO

Direct Score Preference Optimization (DSPO) Zhu et al. is a fine-tuning method for text-to-image (T2I) diffusion models that aligns generation with human preferences. Unlike training-free or proxy-based approaches, DSPO optimizes the model directly using a preference-based objective.

Given two images $\mathbf{x}_t$ (preferred) and $\mathbf{x}'_t$ (less preferred) under condition $\mathbf{c}$, the goal is to align the denoising model's score predictions with human preferences.

We define the following terms:

$$\mathcal{L}_{\text{pred}} = \|\epsilon_{\theta,t+1} - \epsilon_{t+1}\|_2^2 \tag{17}$$

$$\mathcal{L}_{\text{ref}} = \|\epsilon_{\theta,t+1} - \epsilon_{\text{ref},t+1}\|_2^2 \tag{18}$$

The final DSPO objective is:

$$\mathcal{L}_{\text{DSPO}} = \mathcal{L}_{\text{pred}} - \lambda \cdot (1 - \sigma(r(\mathbf{c}, \mathbf{x}_t) - r(\mathbf{c}, \mathbf{x}'_t))) \cdot \mathcal{L}_{\text{ref}} \tag{19}$$

This encourages the model to match the target denoising behavior while differentiating between preferred and less preferred generations. λ is a hyperparameter controlling the strength of the preference-based supervision.

B A Summary of Notations Used in Our Approach

In this section, we summarize the key symbols and terms used throughout the paper. These notations describe the main components of our formulation, including prompts, images, perturbations, masks, and the structure of training data (see Table 2).

Table 2: Summary of notation used in our PREFINE framework.

Symbol	Description
$\mathcal{T}$	Dataset
$\mathcal{T}'$	Expanded Dataset
P	Text prompt
I_w	Winning (reference) image
I_l	Losing (perturbed) image
$\{I_1, I_2, \ldots, I_N\}$	Set of perturbed candidates
$\mathcal{D}(\cdot)$	Visual perturbation function
$\tilde{I}$	Perturbed image produced by $\mathcal{D}$
Ω	Masked region used for distortion
R	Random rectangular subregion
M	Blending mask (binary or soft)
α	Soft blending coefficient
$f(\cdot)$	Reward function used for scoring
s_i	Reward score assigned to image I_i
$B_1, B_2, \ldots, B_r$	Difficulty-based curriculum bins
$k_1, k_2, \ldots, k_r$	Number of samples selected per bin
$\mathcal{B}$	Full set of training triplets (I_w, I_i, s_i)
$\mathcal{B}'$	Diversity-aware selected subset from $\mathcal{B}$
N	Total number of candidate distortions
H, W	Image height and width

C PREFERENCE EXPANSION WITH VISUAL PERTURBATION

To better understand the limitations of model-generated preference data, we visualize and compare outputs generated by the base model against those produced using our PREFINE framework (Figure 6). While the base model's outputs often lack meaningful variation and consistent supervision signals, PREFINE introduces structured perturbations that are both semantically and visually diverse. These perturbations serve as stronger supervisory signals, enabling the model to learn fine-grained preferences more effectively.

C.1 PERTURBATION CATEGORIZATION

We apply a total of 15 perturbations, grouped into two categories: global (9) and local (6) distortions. Table 3 summarizes each perturbation along with a brief description of its visual effect.

D DIVERSITY SAMPLING

D.1 KNAPSACK ANALOGY FOR CORESET SAMPLING

We draw a formal connection between our Coreset sampling strategy and the classical *0/1 Knapsack problem*.

Standard 0/1 Knapsack Formulation. Given a set of n items, each with an associated weight $w_i \in \mathbb{R}_{>0}$ and value $v_i \in \mathbb{R}$, and a knapsack of capacity W, the objective is to select a subset $S \subseteq \{1, \ldots, n\}$ that maximizes total value under the weight constraint:

$$\max_{S \subseteq \{1,\ldots,n\}} \sum_{i \in S} v_i \quad \text{subject to} \quad \sum_{i \in S} w_i \leq W, \quad x_i \in \{0, 1\} \tag{20}$$

Mapping to Our Problem. In our case, we aim to select $M/3$ triplets from each difficulty bucket $(\mathcal{B}_E, \mathcal{B}_M, \mathcal{B}_H)$ such that the selected subset is maximally diverse in terms of difficulty scores. Each triplet $(I_w, I_i, s_i) \in \mathcal{B}$ is treated as an item, with:

Table 3: List of perturbations used to simulate visual degradation, organized into global and local operations based on their spatial effect. Each entry includes representative hyperparameter settings used to control the strength or style of distortion.

Perturbation	**Description**	**Hyperparameters (Example Values)**
Global Perturbations		
Gaussian Blur	Smooths the image using a random Gaussian kernel, simulating loss of fine detail.	Kernel size: odd $\in$ [3, 13]
Gaussian Noise	Adds mild random Gaussian noise to mimic texture artifacts or sensor-like noise.	Std. dev. $\sim \mathcal{U}(5, 40)$
Salt-and-Pepper Noise	Introduces sparse white and black pixels, imitating binary pixel errors.	Amount $\sim \mathcal{U}(0.002, 0.05)$
Channel Swapping	Randomly permutes or drops RGB channels, introducing unnatural color compositions.	Action $\in$ {swap, drop, swap_bw}
Shearing	Applies affine shear to simulate mild geometric warping in horizontal and vertical directions.	Shear factor in X and Y $\in \mathcal{U}(-0.25, 0.25)$
Posterization	Reduces color depth per channel, producing flat, banded color regions.	Bits $\in$ [1, 6] per channel
Elastic Transformation	Applies smoothed random displacement fields, causing flexible geometric deformations.	$\alpha \in [30, 80]$, $\sigma \propto$ image size
JPEG Compression	Emulates medium-quality JPEG artifacts due to lossy encoding and blocking.	Quality $\in$ [1, 40]
Localized Unmasked Perturbations		
Color Jitter	Randomly alters brightness and contrast, simulating lighting or post-processing inconsistencies.	Contrast $\alpha \sim \mathcal{U}(0.8, 1.6)$, Brightness $\beta \sim \mathcal{U}(-20, 20)$
Pixelation	Downsamples and upsamples random regions to simulate blocky, low-resolution textures.	Block size: $10\% \sim 30\%$, Pixel size $\in$ [4, 20]
Erase with Inpainting	Random regions are masked and filled using nearby content, mimicking over-smooth reconstruction.	Num regions $\in$ [1, 3], Shape $\in$ {circle, rectangle}
Localized masked Perturbations		
Swirl Perturbation	Applies local rotational warping around a random center, resembling twisted spatial patterns.	Strength $\in [10, 20]$, Radius $\in [100, 300]$
Twist (Polar Rotation)	Applies center-weighted angular perturbation, creating spiraling or warped structures.	Strength = 5.0 (fixed)
Radial Zoom	Simulates zoom-like radial expansion centered in the image, exaggerating scale and geometry locally.	Factor = 0.001 (fixed)
Sine Wave Perturbation	Displaces pixels using sinusoidal horizontal shifts, imitating spatial wobble or ripple artifacts.	Amplitude = 20, Wavelength = 50

Table 4: Training configurations for SD v1.5 and SDXL-Base Models Using DPO and DSPO.

(a) SD v1.5 with DPO

Parameter	Value
Model	stable-diffusion-v1-5
VAE	–
Mixed Precision	fp16
Resolution	512
Batch Size	16
Grad Accumulation	2
8bit Adam	True
Rank	8
Learning Rate	1e-8
Scale LR	True
LR Scheduler	constant_with_warmup
Warmup Steps	500
Max Steps	350
Checkpoint Steps	50
Beta DPO	0.001

(b) SDXL-Base with DPO

Parameter	Value
Model	stabilityai/sdxl-base-1.0
VAE	madebyollin/sdxl-vae-fp16-fix
Mixed Precision	fp16
Resolution	1024
Batch Size	16
Grad Accumulation	2
8bit Adam	True
Rank	8
Learning Rate	1e-8
Scale LR	True
LR Scheduler	constant_with_warmup
Warmup Steps	200
Max Steps	350
Checkpoint Steps	100
Beta DPO	5000

(c) SD v1.5 with DSPO

Parameter	Value
Model	stable-diffusion-v1-5
VAE	madebyollin/sdxl-vae-fp16-fix
Resolution	512
Batch Size	16
Grad Accumulation	2
8bit Adam	True
Rank	8
Learning Rate	5e-5
LR Scheduler	constant
Warmup Steps	0
Max Steps	1000
Checkpoint Steps	25

(d) SDXL-Base with DSPO

Parameter	Value
Model	stabilityai/sdxl-base-1.0
VAE	madebyollin/sdxl-vae-fp16-fix
Resolution	1024
Batch Size	16
Grad Accumulation	2
8bit Adam	True
Rank	8
Learning Rate	5e-5
LR Scheduler	constant
Warmup Steps	0
Max Steps	1000
Checkpoint Steps	25

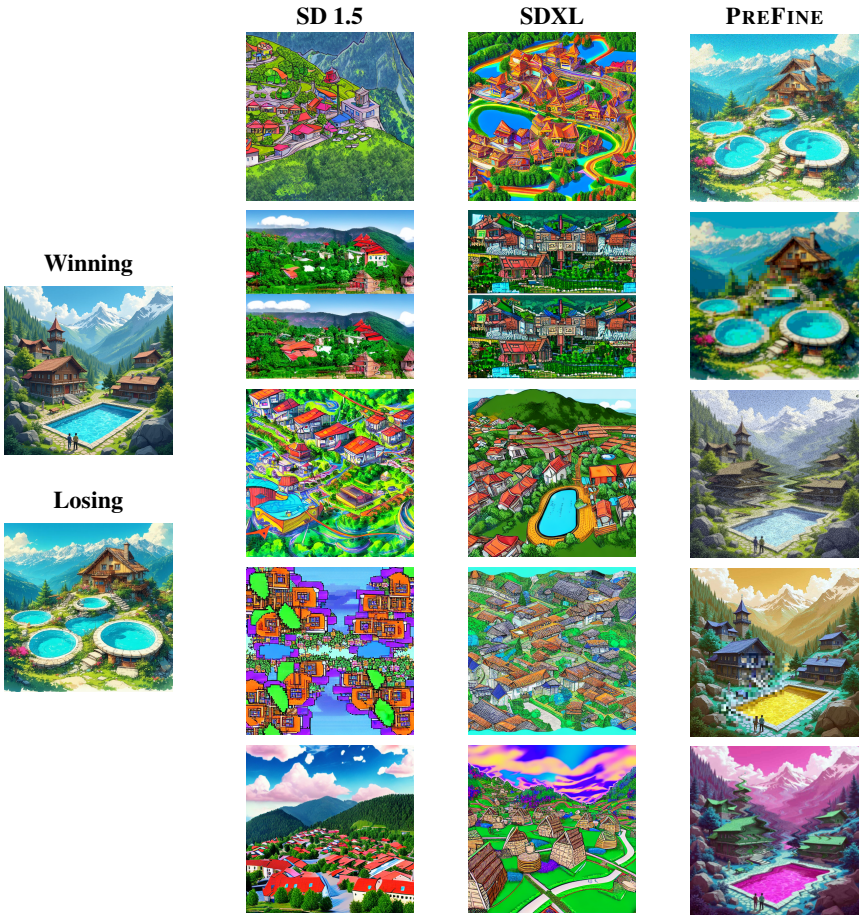


Figure 6: Losing candidates synthesized with the pre-trained models and PREFINE. We observe poor prompt fidelity and loss of information in the images synthesized with the pre-trained models. To synthesize the image with the pre-trained models, we extend the original prompt (*"Generate an image of a Blockchain village in Carpathian mountains with 2 pools, vibrant manga style, dynamic angles, expressive characters..."*) with the phrase *"...with the following perturbations: [perturbations]"*. The perturbation sets used are:**(A)**: salt-and-pepper noise, color jitter, Gaussian blur, Gaussian noise, radial zoom; **(B)**: posterization, Gaussian blur, posterization, pixelation, Gaussian blur, pixelation; **(C)**: Gaussian blur, sine wave distortion, channel swap, elastic transformation, sine wave distortion, salt-and-pepper noise; **(D)**: JPEG compression, pixelation, channel swap; **(E)**: sine wave distortion, Gaussian blur, JPEG compression, channel swap.

- **Unit weight**: $w_i = 1$, and a hard constraint of $M/3$ items per bucket (analogous to knapsack capacity).
- **Dynamic value**: The "value" of including a triplet is not fixed, but instead depends on its score difference with other selected items.

Formally, the objective becomes:

$$\mathcal{B}'_k = \underset{|\mathcal{B}'_k|=M/3}{\arg\max} \sum_{j=1}^{M/3-1} |s_{j+1} - s_j| \tag{21}$$

This reward structure is *state-dependent*, as the inclusion value of a new item depends on the composition of the current selection. This transforms the problem into a knapsack variant with dependent or interaction-based item values.

Table 5: Per-method best scores (bolded) across batch sizes (4, 8, 16) for each dataset using (a) SD-v1.5 and (b) SDXL-Base. Each method is compared independently.

(a) SD-v1.5

Dataset	Method	BS	Aes.	ImgReward
Pick-a-Pic V2	Ours + Diff.-DPO	4	58.6	**63.0**
		8	**63.4**	62.0
		16	57.6	62.2
	Ours + DSPO	4	49.6	53.0
		8	47.0	50.0
		16	**47.4**	**53.0**
PartiPrompt	Ours + Diff.-DPO	4	**65.8**	58.6
		8	62.8	59.8
		16	59.4	**61.2**
	Ours + DSPO	4	51.8	**52.6**
		8	54.2	49.6
		16	**55.2**	50.6
GenAIBench	Ours + Diff.-DPO	4	62.6	51.8
		8	**65.2**	**53.8**
		16	60.0	54.2
	Ours + DSPO	4	52.6	**52.8**
		8	**54.0**	48.8
		16	50.8	48.4

(b) SDXL-Base

Dataset	Method	BS	Aes.	ImgReward
Pick-a-Pic V2	Ours + Diff.-DPO	4	56.4	62.0
		8	54.6	62.6
		16	**54.2**	**64.8**
	Ours + DSPO	4	50.6	50.2
		8	46.4	**52.6**
		16	**49.4**	51.2
PartiPrompt	Ours + Diff.-DPO	4	**55.8**	56.4
		8	52.0	56.2
		16	55.6	**58.0**
	Ours + DSPO	4	47.6	47.8
		8	**49.0**	49.2
		16	49.0	**49.8**
GenAIBench	Ours + Diff.-DPO	4	**57.6**	57.4
		8	57.0	57.6
		16	57.8	**59.4**
	Ours + DSPO	4	48.0	46.8
		8	49.2	47.6
		16	**50.4**	**49.0**

E Experimental Details

E.1 Hyperparameters

Table 4 provides the training configurations used for various scripts involving SD1.5 and SDXL-Base models with either DPO or DSPO. It outlines key hyperparameters such as model type, resolution, optimizer settings, learning rate schedules, and checkpoint intervals, serving as a reference for reproducibility and setup transparency.

F Additional Results

F.1 Quantitative Results

To assess whether our improvements stem primarily from data augmentation effects (e.g., increased batch size or sampling diversity), we conduct an analysis across different batch sizes (M=4, 8, 16) during preference optimization (Table 5). Interestingly, the results do not exhibit a consistent trend where larger batch sizes lead to better performance. In several cases, smaller batches outperform larger ones—for instance, in SD v1.5, on the Pick-a-Pic V2 dataset with Ours + DSPO, the best ImageReward score is achieved with M=8 rather than M=16. Similarly, on GenAIBench with Ours + Diffusion-DPO, ImageReward peaks at M=16. These findings suggest that the gains observed with our method are not merely a result of increased data exposure through larger batches. Instead, they underscore the robustness and generalization ability of our optimization framework across a range of training configurations, independent of sampling scale.

F.2 Qualitative Results

We provide additional qualitative comparisons in Figure 7, Figure 8, and Figure 9. These visual examples compare outputs from SD v1.5 and SDXL-Base models fine-tuned using standard DPO and our proposed DSPO approach, with and without using the PreFine dataset. Across a range of challenging prompts, our method produces outputs with noticeably better prompt adherence, clearer compositional structure, and improved visual quality. These comparisons highlight how PreFine enhances preference alignment during training, resulting in more coherent and aesthetically pleasing generations.

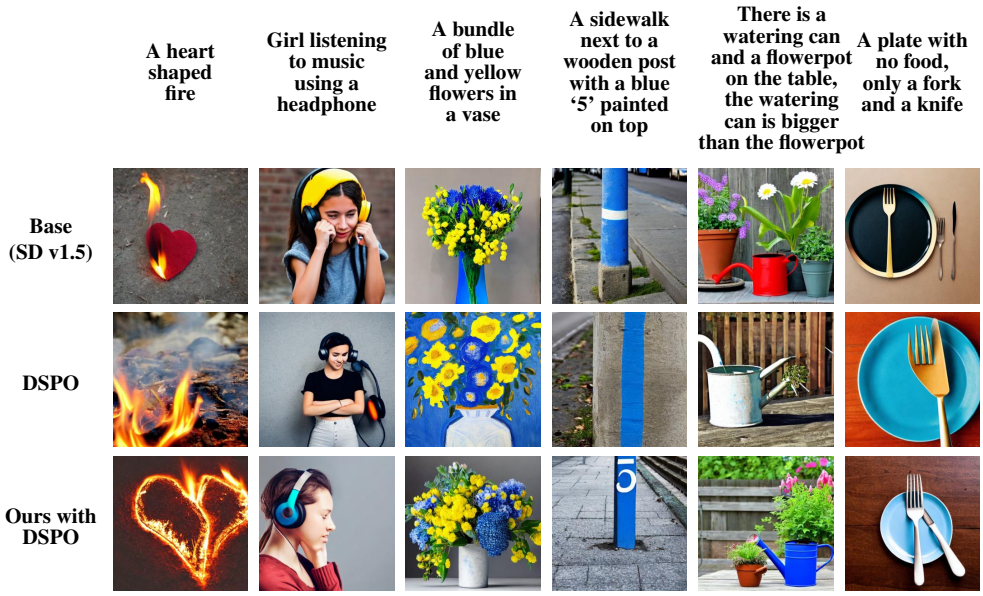


Figure 7: Qualitative performance of preference alignment of SD v1.5 with PREFINE using DSPO. We present two prompts each from Pick-a-Pic V2, PartiPrompt, and GenAIBench respectively. The fine-grained details in the generated images and the aesthetic quality have improved significantly with PREFINE.

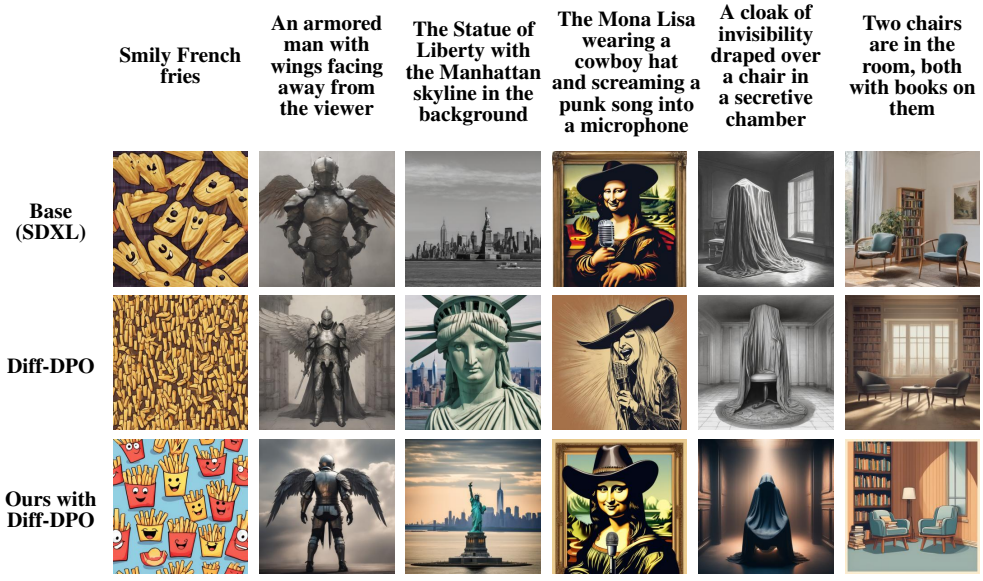


Figure 8: Qualitative performance of preference alignment of SDXL-Base with PREFINE using Diffusion-DPO. We present two prompts each from Pick-a-Pic V2, PartiPrompt, and GenAIBench respectively. The fine-grained details in the generated images and the aesthetic quality have improved significantly with PREFINE.

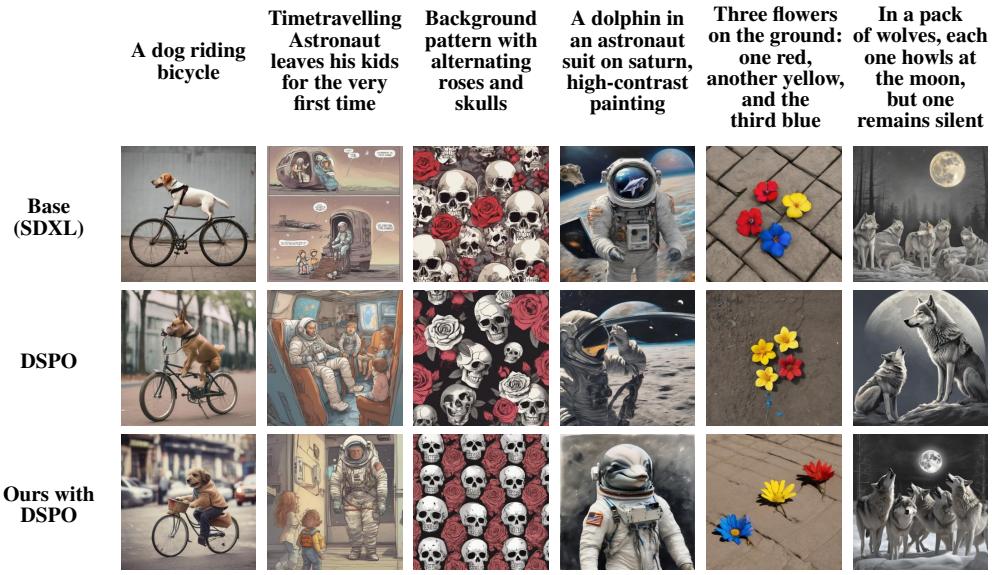


Figure 9: Qualitative performance of preference alignment of SDXL-Base with PREFINE using DSPO. We present two prompts each from Pick-a-Pic V2, PartiPrompt, and GenAIBench respectively. The fine-grained details in the generated images and the aesthetic quality have improved significantly with PREFINE.