Aligning Text-to-Image Diffusion Models using Human Utility Optimization and Low-Rank Adaptation

Wendy Yin
Stanford University
Department of Economics
wendyyin@stanford.edu

Yicheng Zhang
Stanford University
Department of Computer Science
yicheng4@stanford.edu

Yiwen Zhang
Stanford University
Department of Computer Science
leonardz@stanford.edu

Abstract

Text-to-image diffusion models like Stable Diffusion are good at generating high-fidelity images but often fail to adhere to specific or niche artistic styles due to the limitation from the broad nature of their pre-training data. This project aims to address this style alignment gap by investigating whether using parameter-efficient fine-tuning techniques and human-feedback objectives can align models to fine-grained artistic preferences. We employ Low-Rank Adaptation (LoRA) on a Stable Diffusion checkpoint and compare several preference-alignment strategies, including Diffusion-DPO, Diffusion-KTO and SPIN-Diffusion. Our results, evaluated on automated metrics like PickScore & CLIP Score, demonstrate that advanced alignment methods significantly outperform baselines. In particular, SPIN-Diffusion achieved the highest human preference score, closely followed by Diffusion-KTO, highlighting the effectiveness of self-play and direct utility optimization. In side-by-side comparisons, the Diffusion-KTO model consistently preserves finer details, such as fur texture, and maintains more vivid, well-saturated colors across both photographic and stylized prompts. These findings suggest that human utility optimization is a promising and efficient pathway for achieving high-fidelity stylistic control in generative models, enabling critical downstream applications in art and design.

1. Introduction

Text-to-image diffusion models such as Stable Diffusion [11] and DALL-E [10] have rapidly become the backbone of contemporary visual-content generation. Their ability to map arbitrary natural-language prompts onto high-fidelity images has unlocked a wide array of applications. Yet, despite impressive breadth, these models remain coarse instruments when users demand adherence to *highly specific* or *niche* artistic styles. Constrained by the heterogeneous—and often mainstream—signal in their pre-training data, they tend to average over stylistic nuances, yielding images that are aesthetically pleasing but misaligned with idiosyncratic tastes. Addressing this style-specific alignment gap is essential for downstream domains such as concept-art prototyping, personalized game asset creation, and cultural-heritage preservation, where fine-grained artistic fidelity is non-negotiable.

This project tackles the challenge of stylistic alignment through targeted, data-efficient fine-tuning. We mainly study if *parameter-efficient* diffusion model can be updated through Low-Rank Adaptation (LoRA) adapters to be aligned to niche, fine-grained artistic preferences *using human-feedback objectives alone*, and investigate whether such alignment translates into measurable gains over both reconstruction-only fine-tuning and existing pairwise-preference baselines. The input to our algorithm is a set of images representing a target artistic style, along with text prompts. We then use a Stable Diffusion model with LoRA adapters, which we fine-tune by directly optimizing for human utility using objectives like Diffusion-DPO, Diffusion-KTO and SPIN-Diffusion. The final output is a model capable of generating novel images that faithfully capture the desired artistic style from new text prompts.

Formally, given (i) a frozen text encoder and U-Net backbone $\mathcal{M}_0$, (ii) a prompt distribution $\mathcal{P}$, and (iii) a preference corpus $\mathcal{D} = \{(p_k, I_k^+, I_k^-)\}_{k=1}^N$ or binary likes $\{(p_k, I_k, y_k)\}_{k=1}^N$, we ask whether there exists a com-

pact parameter set $\theta^\star$ (LoRA rank $r \ll d$) such that the adapted sampler $\mathcal{M}_{\theta^\star}$ maximizes expected human utility $\mathbb{E}_{p\sim\mathcal{P}}\big[U_{\text{human}}(\mathcal{M}_{\theta^\star}(p))\big]$ subject to a tight complexity budget, and how its performance compares against (a) the un-adapted $\mathcal{M}_0$, (b) supervised DreamBooth-style reconstructions, and (c) state-of-the-art preference-alignment methods like Diffusion-DPO. This framing unifies our empirical study across binary, pairwise, and self-play objectives while isolating the value of LoRA-based updates for stylistic fidelity.

In this work, we navigate these interconnected domains by specifically focusing on:

- **Simplified Data Collection and Utility:** We explore the efficacy of Kahneman-Tversky Optimization (KTO), which promises robust alignment using only binary feedback (e.g., likes/dislikes from a preference corpus like the Laion Art subset with its aesthetic scores). This potentially streamlines the often costly and complex **data collection** phase (axis i) compared to methods requiring explicit pairwise comparisons.
- **Advanced Optimization Objectives:** We systematically compare Diffusion-KTO against other state-of-the-art **optimization objectives** like Diffusion-DPO and a simpler BCE-based preference loss, providing insights into their relative strengths for fine-grained stylistic control.
- **Parameter and Data Efficiency:** Our entire investigation is grounded in **data-efficiency techniques** (axis iii), primarily Low-Rank Adaptation (LoRA). This not only makes our approach computationally tractable but also specifically tests the hypothesis that significant stylistic alignment can be achieved with minimal parameter updates and focused preference data.

By focusing on the intersection of human utility optimization and parameter-efficient tuning, we aim to demonstrate a practical path towards achieving nuanced artistic control in large-scale diffusion models.

2. Related Work

Recent progress in aligning text-to-image models with human preferences can be understood across three interconnected domains: the creation of preference datasets, the development of optimization objectives to leverage this data, and the invention of techniques to improve data efficiency. Our work is situated at the intersection of these domains, using parameter-efficient methods and human utility optimization to align models with niche artistic styles.

Preference Datasets for Text-to-Image Alignment. The foundation of preference alignment lies in the data used to represent human aesthetic judgments. Pairwise preference datasets have become a popular standard. Pick-a-Pic introduced a public corpus of crowd-sourced pairwise votes, enabling systematic comparison of model outputs [5]. ImageRewardDB extended this idea, gathering 137k expert comparisons and distilling them into a CLIP-based reward model [19]. Human Preference Score v2 (HPSv2) further scaled to 800 k comparisons and established a robust automatic metric [17].
Recognizing that a single preference score can be limiting, researchers have developed datasets with more granular feedback. VisionReward, for example, decomposed user judgments into interpretable sub-scores, furnishing multi-attribute labels for both image and video generation [18]. And there are more specialized datasets where we describe in the Dataset section.

Optimization Objectives for Alignment. Given these datasets, various optimization objectives have been proposed to align diffusion models. Early Reward-model pipelines, such as ReFL, tune generators directly against the ImageReward scorer [19]. Direct Preference Optimization (DPO) [9], adapted from language models, has become a state-of-the-art technique. Diffusion-DPO adapts Direct Preference Optimization to diffusion likelihoods, achieving state-of-the-art appeal on SDXL without explicit reinforcement learning [16]. D3PO further reduces memory overhead by operating in the denoising latent space [20]. Our work heavily leverages a successor to these methods, Diffusion-KTO. Kahneman-Tversky Optimization (KTO) [2] aims to improve the efficiency and quality of LLM alignment while reducing the need for expensive preference data. KTO represents a significant advancement by eliminating the need for pairwise data entirely. Based on KTO, Diffusion-KTO offers per-sample utility calculation and thus it can maximize expected human utility using only binary feedback (e.g., likes/dislikes), which dramatically simplifies the data collection process. [6].

Data-Efficient Alignment Techniques. SPIN-Diffusion employs a self-play strategy where the current model is compared against a frozen, earlier checkpoint to generate synthetic preference pairs. This allows the model to bootstrap its own alignment signal, effectively reducing the need for human data [22]. Moreover, FiFA proposes automated filtering that can accelerate DPO training by two orders of magnitude, making the alignment process faster and more efficient [21].

Connection to historical development of utility function. Under the von-Neumann–Morgenstern axioms, binary feedback or pairwise comparisons can be em-

bedded in a cardinal utility function whose expectation is the object of optimisation. Random-utility theory interprets each observed vote as a noisy realisation of latent utility and motivates the logistic losses used in DPO and KTO. Indeed, Diffusion-DPO's objective is formally identical to maximum-likelihood estimation in McFadden's conditional-logit model [7]. Moreover, Afriat's revealed-preference theorem guarantees that, absent preference cycles, a continuous, monotone utility rationalizes any finite set of binary choices [1]; this justifies the internal-consistency checks commonly applied to feedback datasets. Viewed through this lens, reward-model pipelines estimate a surrogate utility index, whereas reward-free methods such as KTO maximize expected utility directly—mirroring the distinction between indirect and direct utility estimation in micro-econometrics.

3. Methods

Our core objective is to align a pre-trained, large-scale diffusion model with fine-grained, niche artistic styles. To do this efficiently, we adopt a parameter-efficient fine-tuning (PEFT) strategy, ensuring that only a small fraction of the model's parameters are updated. This allows for rapid experimentation and makes the stylistic adaptation of massive models computationally tractable. All our experiments start from the same `Stable Diffusion v1.5` checkpoint. We then compare a supervised reconstruction-based baseline against several advanced human-preference alignment algorithms, all implemented within a unified LoRA framework. With this approach, we can clearly measure the benefits of using human feedback and compare the results with traditionally fine-tuned models.

We plan to compare four preference-alignment strategies under a common *parameter-efficient* setting: all methods start from the same `Stable Diffusion v1.5` checkpoint and update only rank-8 LoRA adapters in the U-Net (3.1). The variants differ in how human feedback enters the optimisation objective (3.2), yielding a clean ablation of preference signals versus reconstruction-only fine-tuning. Our evaluation includes automatic metrics (PickScore, CLIP Score).

3.1. Minimal Baseline: Binary-Preference LoRA

3.1.1 Parameter-Efficient Fine-Tuning with LoRA

LoRA is a technique that enables the efficient fine-tuning of large models by injecting trainable, low-rank matrices into the model's architecture while keeping the original pre-trained weights frozen. In the attention blocks of the U-net, we would augment each weight matrix $W_0 \in \mathbb{R}^{d\times d}$ as follows:

$$W_\theta \;=\; W_0 + A\,B^\top, \qquad A\in\mathbb{R}^{d\times r},\; B\in\mathbb{R}^{d\times r},\; r\ll d, \tag{1}$$

This decomposition helps reduce the number of trainable parameters for the layer from d^2 to only $2dr$ [4].
In our project, we utilize a rank of $r = 8$ for all LoRA adapters in order to obtain a balance between model expressiveness and parameter efficiency. This approach significantly reduces the memory and computational requirements for fine-tuning while demonstrating strong performance in adapting the model's behavior.

3.1.2 Diffusion Model Preliminaries

Our approach is built upon the framework of latent diffusion models. The process begins by encoding a training image into a lower-dimensional latent representation, $\mathbf{z}_0$, using a pre-trained Variational Autoencoder (VAE). The forward diffusion process then gradually adds Gaussian noise to this latent over a series of timesteps t. Following the Denoising Diffusion Implicit Models (DDIM) formulation [14], the noisy latent $\mathbf{z}_t$ at any timestep t can be sampled as:

$$\mathbf{z}_t \;=\; \sqrt{\bar\alpha_t}\,\mathbf{z}_0 + \sqrt{1-\bar\alpha_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}\sim\mathcal{N}(0,I), \tag{2}$$

Here, $\boldsymbol{\epsilon}$ is the noise samples from a standard normal distribution $\mathcal{N}(0, I)$, and $\bar\alpha_t$ is a pre-defined noise schedule parameter that controls the signal-to-noise ratio at timestep t. The objective of the denoising model is to predict the noise $\boldsymbol{\epsilon}$ that was added to the latent, given the noisy latent $\mathbf{z}_t$ and a conditioning input text prompt. We apply LoRA adapters to the cross-attention layers of the denoising model to guide the generation process.

3.1.3 Binary-Preference LoRA

Our minimal baseline preference-based model uses a straightforward binary cross-entropy (BCE) loss. For each image in our preference dataset, which is labeled as $y = 1$ for *exclusive_win* and 0 otherwise, the LoRA-augmented U-Net predicts $\hat{\boldsymbol{\epsilon}}_\theta$. The preference loss is formulated as:

$$\mathcal{L}_{\text{pref}} = \mathrm{BCE}\big(-\mathrm{MSE}(\hat{\boldsymbol{\epsilon}}_\theta, \boldsymbol{\epsilon}),\, y\big), \tag{3}$$

In this objective, the negated pixel-wise Mean Squared Error (MSE) between the predicted noise $\hat{\boldsymbol{\epsilon}}_\theta$ and the true noise $\boldsymbol{\epsilon}$ is used as a logit. This construction, echoing the one used for the ImageReward model [19], effectively trains the model to produce a lower reconstruction error on preferred images and a higher error on disliked images, thus implicitly learning the preference distribution.

3.2. Enhanced Alignment Variants

(i) DreamBooth+LoRA. As a *supervised* baseline to benchmark against non-preference methods, we fine-tune via the reconstruction loss of DreamBooth, a popular technique for subject-driven generation. DreamBooth fine-tunes

a text-to-image model on a few images of a specific subject or style and its objective is to accurately recreate the VAE latents of the provided images when conditioned on a unique identifier prompt without any human preference data [12]. Its could help us quantify the performance gains achievable through human utility optimization.

(ii) Diffusion-DPO. Direct Preference Optimization (DPO) is a powerful and stable method for aligning models with human preferences that bypasses the need for an explicit reward model [9]. Adapted for diffusion models, Diffusion-DPO learns directly from a dataset of preference pairs, where each entry consists of a prompt. For each prompt we sample a "winner" image $\mathbf{x}^+$ and "loser" $\mathbf{x}^-$. With a temperature β, DPO minimizes

$$\mathcal{L}_{\text{DPO}} \;=\; -\log \sigma\big(\beta\,[r_\theta(\mathbf{x}^+) - r_\theta(\mathbf{x}^-)]\big), \qquad (4)$$

where r_θ is the per-image implicit reward function parameterized by the diffusion model itself [16]. The loss works by maximizing the margin between the implicit reward of the winner image and the loser image. The temperature parameter β controls how strongly the loss penalizes the model for mismatching the pair, with higher values of β leading to a stronger level of preference enforcement.

(iii) Diffusion-KTO. Kahneman-Tversky Optimization (KTO) further simplifies the data requirements for preference alignment. Unlike DPO, KTO dispenses with pairwise comparisons and can learn directly from binary labels. The objective of KTO is to maximize the expected utility of the images generated by the model from binary likes:

$$\max_\theta \; J(\theta) = \mathbb{E}_{\mathbf{z}\sim p_\theta}\big[u(\mathbf{z})\big] \qquad (5)$$

where p_θ is the image distribution and $u(\mathbf{z})$ is a utility function derived from the binary human feedback. Since the expectation is relatively hard to control, KTO uses a score-function estimator with baseline λ to compute the policy gradient:

$$\nabla_\theta J = \mathbb{E}\big[\nabla_\theta \log p_\theta(\mathbf{z})\big(u(\mathbf{z}) - \lambda\big)\big] \qquad (6)$$

In the estimator, λ serves as a baseline to reduce the variance of the gradient estimates, leading to more stable training [6]. KTO's ability to learn from simple, unpaired human feedback makes it a highly data-efficient and more flexible compared with other alignment methods.

(iv) SPIN-Diffusion. Self-Play fIne-tuNing (`SPIN`) is a technique designed to reduce the need for large quantities of human-annotated data by having the model generate its own training signals. In SPIN-Diffusion, the model generates synthetic pairs by comparing the current model to a frozen copy θ^- and applying the DPO loss (4) to those pairs. This process creates a curriculum where the model continuously refines its own notion of model quality, bootstraps its alignment and discovers hard negative examples without additional human annotation. In this way SPIN-Diffusion could help us half the data requirement [22].

3.3. Hypotheses

Building on the distinct characteristics of the alignment strategies and our goal of achieving nuanced stylistic control via parameter-efficient means, we hypothesize (H1) that DPO and KTO outperform the reconstruction-only DreamBooth baseline, (H2) that KTO matches DPO despite needing only unpaired likes, and (H3) that SPIN yields further gains by self-generating hard negatives.

4. Dataset and Features

The success of preference alignment is highly dependent on the quality and nature of the underlying dataset. Human–feedback corpora for text–to–image alignment now fall into three principal categories. *Binary preference sets* such as ImageRewardDB (136k prompt–image pairs) [19] and Human Preference Dataset v2 (HPS v2; 800 k pairs) [17] provide high–signal supervision for relative-quality objectives but are largely biased toward photographic prompts. *Continuous–score resources* exemplified by the 12 M–image LAION–Aesthetics collection [13] offer web-scale coverage, yet their scalar labels are unsuited to pairwise–difference losses. Finally, *free-form critique corpora*, notably the Reddit Photo Critique Dataset (RPCD; 74 k images, 220 k comments) [15], deliver nuanced textual assessments but remain too small and noisy for primary training.

For this project we adopt the **Laion Art subset** as our core training source [3]. Curated for illustrative and fantastical content, it aligns perfectly with our goal of training for niche artistic styles. For our experiments, we utilized a filtered subset of high-resolution examples for training. A key advantage of the Laion Art subset is its uniform 512×512 resolution, which streamlines our data pipeline as it eliminates the need for resizing. No data augmentation techniques, such as random flips or crops, were applied in our experiments, as our main focus is on learning from the specific composition and details of the provided artistic examples. To extract image features from the dataset, we use the a pre-trained Variational Autoencoder (VAE) to encode the images into a low-dimensional latent space.

To train our preference-based models, we leveraged the dataset's built-in aesthetic scalar score for each image. These scores were converted into deterministic *exclusive_win/lose* labels, allowing us to generate the binary feedback required for the Diffusion-KTO and baseline

models without manual annotation, preserving sample efficiency.

5. Experiments, Results, and Discussion

Our experimental evaluation aims to quantify the effectiveness of different preference-alignment strategies in enhancing the stylistic fidelity of text-to-image diffusion models. We compare our proposed methods—Diffusion-KTO, Diffusion-DPO, and SPIN-Diffusion, all leveraging LoRA for parameter-efficient fine-tuning—against the baseline Stable Diffusion v1.5 (Base SD) and a minimal Binary Cross-Entropy LoRA fine-tuned model (BCE LoRA).

5.1. Experimental Setup

5.1.1 Training Hyperparameters

All fine-tuning variants commenced from the same `Stable Diffusion v1.5` checkpoint and exclusively updated rank-8 LoRA adapters in the U-Net, along with text encoder bias terms. We employed the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight-decay $= 0.01$), a standard choice for its effectiveness in training large neural networks, with weight decay providing regularization. A constant learning rate of 1×10^{-4} was used for LoRA weights and 1×10^{-6} for the text encoder biases, selected based on common practices for LoRA fine-tuning that allow for effective adaptation without destabilizing the pre-trained model. A warm-up phase of 500 steps was used to stabilize initial training. All models were trained for 10,000 steps. Due to computational constraints and the scale of the models, an exhaustive hyperparameter search using extensive cross-validation was not performed for this phase of the project; the chosen values are based on literature recommendations and preliminary experiments aiming for stable and effective training.

Our preliminary experiments for hyperparameter refinement, while not constituting an exhaustive search, were crucial for ensuring stable and effective training across all compared methods. We initiated with hyperparameters such as learning rates (LoRA: 5×10^{-4}, text encoder biases: 1×10^{-5}), AdamW optimizer settings ($\beta_1 = 0.9, \beta_2 = 0.999$, weight-decay $= 0.05$), and warm-up steps (100), selecting values well-established in LoRA and diffusion model fine-tuning literature. To validate these choices and to tune method-specific parameters, such as the β temperature in Diffusion-DPO (Equation 4) and analogous sensitivity points in Diffusion-KTO, we conducted short trial runs for each alignment strategy. These typically involved training for approximately 10-20% of the total 10,000 steps on a random subset of the Laion Art dataset. During these trials, we primarily monitored the stability of the training loss curves and qualitatively assessed image outputs generated from a fixed set of diverse prompts. This allowed us to check for coherent stylistic application, semantic integrity, and the absence of common training pathologies like mode collapse or excessive artifacts. For instance, for Diffusion-DPO's β, we explored a small set of values (e.g., 0.8, 0.9, 0.99, 1.0) as guided by prior work, selecting the one that demonstrated a good balance between effective preference differentiation and stable learning dynamics in these initial outputs. This pragmatic tuning process aimed to establish a robust and equitable hyperparameter baseline for all compared methods within our computational constraints, rather than to individually optimize each method to its theoretical peak performance.

Mini-batch sizes were optimized to maximize GPU utilization on a single A100-80GB GPU, typically ranging from 4 to 8 samples per device depending on the specific memory footprint of the variant. Further details on minor hyperparameter tuning are deferred to supplemental material.

5.1.2 Evaluation Metrics

To assess model performance, we utilized two primary automated metrics prevalent in recent text-to-image generation literature:

1. CLIP Score: This metric measures the semantic similarity between a generated image and its corresponding text prompt. It is calculated as the cosine similarity between the image embeddings and text embeddings produced by a pre-trained CLIP model (Contrastive Language-Image Pre-training) [8]. Given an image I and a text prompt T, let $E_I(I)$ be the image embedding and $E_T(T)$ be the text embedding from CLIP. The CLIP Score is:

$$\text{CLIP Score} = \cos(E_I(I), E_T(T)) = \frac{E_I(I) \cdot E_T(T)}{\|E_I(I)\| \|E_T(T)\|}$$

Higher CLIP scores indicate better alignment between the image content and the textual description.

2. PickScore: This metric [5] is a learned reward model trained on a large dataset of human preferences between pairs of images generated from the same prompt. It aims to predict which image a human would prefer, reflecting aspects like aesthetic quality, prompt adherence, and overall appeal. A higher PickScore suggests that the generated image is more likely to be preferred by humans. Since PickScore is itself a neural network, there isn't a simple equation, but it outputs a scalar value indicating preference.

These metrics were chosen to provide complementary insights: CLIP Score focuses on semantic fidelity to the prompt, while PickScore offers a proxy for human-perceived quality and stylistic preference alignment.

5.2. Quantitative Evaluation

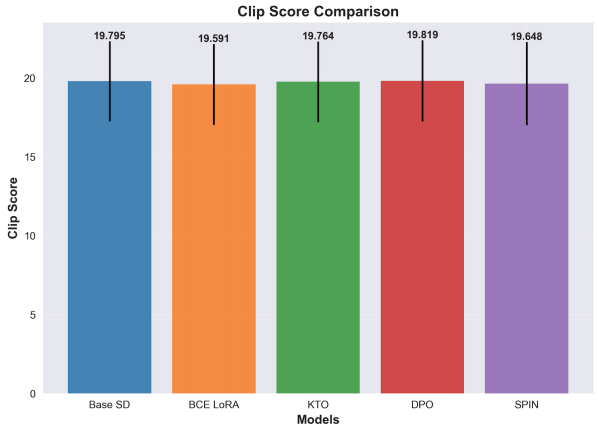
The mean and standard deviation for PickScore and CLIP Score across all evaluated models are summarized in

Table 1 and Figure 1 and 2.

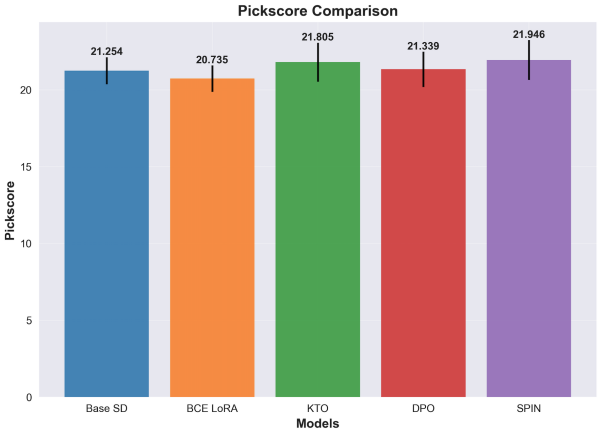
The models were evaluated on a diverse set of prompts, exemplified in Appendix 7.1 (general photographic) and Appendix 7.2 (stylized counterparts). This set was designed to cover varied subjects and styles, allowing for assessment of both general artistic rendering and adherence to specific stylistic keywords (e.g., "watercolor painting," "impressionist style").

Model	PickScore	CLIP Score
Base SD v1.5	21.25 (±0.88)	19.80 (±2.54)
BCE LoRA	20.74 (±0.86)	19.59 (±2.56)
Diffusion-DPO	21.34 (±1.14)	19.82 (±2.57)
Diffusion-KTO	21.80 (±1.27)	19.76 (±2.55)
SPIN-Diffusion	21.95 (±1.30)	19.65 (±2.63)

Table 1: Quantitative Comparison of Different Models.

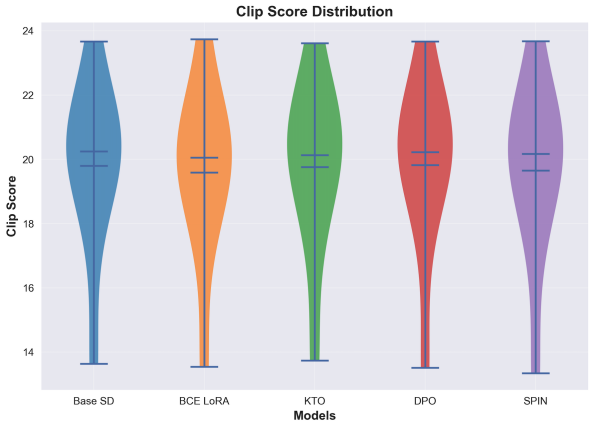


(a) CLIP Score Comparison

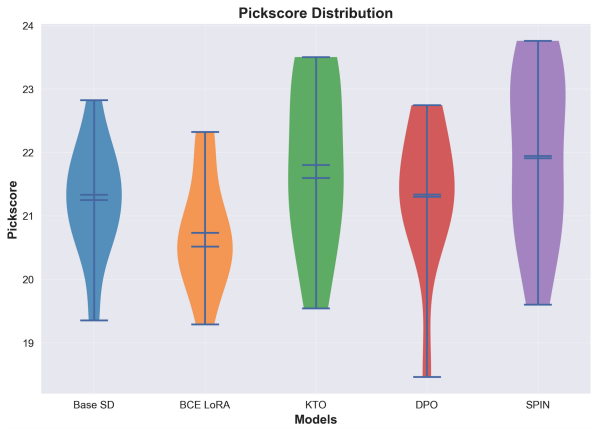


(b) PickScore Comparison

Figure 1: Side-by-side comparison of PickScore and CLIP Score



(a) CLIP Score Distribution



(b) PickScore Distribution

Figure 2: Side-by-side comparison of PickScore and CLIP Score distributions

5.2.1 PickScore Analysis

The PickScore results indicate significant differences in human preference alignment across the models. Notably, SPIN-Diffusion achieved the highest mean PickScore (21.95±1.30), closely followed by Diffusion-KTO (21.80±1.27) These scores represent a substantial improvement over the un-adapted Base SD v1.5 (21.25±0.88). Diffusion-DPO also showed a slight improvement over the Base SD (21.34±1.14). Interestingly, the minimal BCE LoRA model (20.74±0.86) performed worse than the Base SD model, suggesting that a naive application of preference learning with a simple BCE loss on reconstruction error logits may not be sufficient or could even be detrimental to perceived quality if not carefully tuned.

5.2.2 CLIP Score Analysis

Regarding CLIP Scores, all models performed comparably, with mean scores clustered around 19.6 to 19.8. Diffusion-

DPO (19.82±2.57) and the Base SD v1.5 (19.80±2.54) achieved marginally higher scores, though the differences between all models are small relative to their standard deviations. This suggests that while the preference-alignment techniques significantly impact the stylistic qualities favored by PickScore, they largely preserve the fundamental text-to-image semantic alignment. The slight variations might indicate that models focusing more on specific stylistic nuances (as encouraged by preference tuning) might sometimes make minor trade-offs in literal semantic interpretation compared to the base model.

5.3. Qualitative Evaluation

As illustrated in Figure 3, the qualitative differences are striking. The Diffusion-KTO model, for instance, consistently demonstrated an enhanced ability to preserve finer details (e.g., fur texture) and maintain more vivid, well-saturated colors across both photographic and stylized prompts compared to the Base SD v1.5 and the BCE LoRA model. The BCE LoRA model often exhibited oversmoothing, particularly in low-contrast regions. These visual assessments corroborate the quantitative PickScore findings, where Diffusion-KTO substantially outperformed these two models.

5.4. Discussion

The results provide valuable insights into the effectiveness of human utility optimization and parameter-efficient fine-tuning for aligning diffusion models to niche artistic preferences.

Our findings directly address the hypotheses outlined in Section 3.4:

H1 (DPO and KTO outperform reconstruction-only DreamBooth baseline): Both Diffusion-KTO (PickScore: 21.80) and Diffusion-DPO (PickScore: 21.34) significantly outperformed the minimal BCE LoRA preference baseline (20.74) and the Base SD v1.5 (21.25). The qualitative improvements shown by KTO (Figure 3) also suggest a marked enhancement in stylistic fidelity, which is the primary goal of DreamBooth-style fine-tuning.

H2 (KTO matches DPO despite needing only unpaired likes): Our results suggest that Diffusion-KTO not only matches but outperforms Diffusion-DPO in terms of PickScore (21.80 for KTO vs. 21.34 for DPO) with the current dataset and experimental setup. This is a significant finding, as KTO's simpler data requirement (binary likes/dislikes) compared to DPO's pairwise preference pairs makes it a more data-efficient and potentially more scalable approach for preference alignment. The comparable CLIP scores indicate this improved preference alignment does not come at a cost to semantic coherence.

H3 (SPIN yields further gains by self-generating hard negatives): This hypothesis is strongly supported by our results. SPIN-Diffusion achieved the highest PickScore (21.95), surpassing both DPO and KTO. This indicates that the self-play strategy, where the model generates its own training signals by comparing against an earlier version of itself, is highly effective in refining alignment and discovering aspects that contribute to preferred image generation without requiring additional human-annotated data beyond the initial preference corpus (if any is used to kickstart the process, or if it's built upon a KTO/DPO-like objective internally).

(a) Normal prompt: *"A photo of a cat with blue eyes"*

(b) Stylised prompt: *"A watercolor painting of a cat with blue eyes"*

(c) Normal prompt: *"A small cottage in the countryside"*

(d) Stylised prompt: *"An oil painting of a cozy cottage in impressionist style"*

Figure 3: Side-by-side comparison of generation results for two scenes under normal and stylised prompts. Each sub-figure itself juxtaposes outputs from the base Stable Diffusion v1.5, a LoRA fine-tuned model, and the Diffusion KTO model.

5.5. Implications of Findings

The superior performance of KTO and SPIN-Diffusion, both of which leverage human utility optimization principles, underscores the potential of these methods for achieving nuanced artistic control. The success of KTO is particularly promising due to its reduced reliance on complex pairwise preference data. The fact that the parameter-efficient LoRA framework enabled these improvements makes these techniques practical for adapting large-scale diffusion models.

The underperformance of the BCE LoRA model highlights that the choice of optimization objective is critical. Simply encouraging lower reconstruction error on "liked" images and higher error on "disliked" images via a BCE loss on MSE logits does not robustly translate to improved stylistic alignment as measured by PickScore, and may even degrade general quality if it leads to overly conservative or biased outputs. More sophisticated objectives like those in DPO and KTO, which directly model preference probabilities or utility, are clearly more effective.

The consistent CLIP scores across models are reassuring, suggesting that the alignment process primarily refines stylistic aspects without catastrophically forgetting core semantic understanding. However, the subtle trade-offs observed warrant further investigation, particularly in scenarios demanding extremely high fidelity to complex prompts.

6. Conclusion and Future Work

This project investigated the alignment of text-to-image diffusion models with fine-grained artistic preferences using parameter-efficient fine-tuning and human utility optimization. By employing Low-Rank Adaptation (LoRA) on a Stable Diffusion v1.5 checkpoint, we compared several preference-alignment strategies, including a baseline BCE LoRA, Diffusion-DPO, Diffusion-KTO, and SPIN-Diffusion. Our quantitative results, primarily driven by PickScore, demonstrate that advanced alignment techniques leveraging human utility optimization principles significantly enhance stylistic fidelity. Notably, SPIN-Diffusion and Diffusion-KTO emerged as the highest-performing methods, substantially improving perceived image quality over the base model and simpler fine-tuning approaches. Diffusion-KTO's success is particularly compelling as it achieves strong results using only binary preference data, simplifying data collection. SPIN-Diffusion's leading performance highlights the efficacy of self-play mechanisms in generating challenging training examples and continuously refining the model. In contrast, the BCE LoRA model underperformed, suggesting that naive preference objectives are insufficient for capturing nuanced stylistic preferences. All methods largely maintained semantic consistency as per their CLIP scores.

The promising performance of Diffusion-KTO and SPIN-Diffusion underscores the value of directly optimizing for human utility and the potential of self-supervised preference generation. We believe these methods worked better due to their more sophisticated modeling of preferences: KTO by directly maximizing expected utility from simpler feedback, and SPIN by creating an internal curriculum of increasingly difficult preference pairs. These approaches are more robust and aligned with the complex nature of aesthetic judgment than the indirect signal provided by the BCE LoRA baseline.

For future work, several further exploration can be made. An immediate priority is to conduct the planned 1,000-sample A/B human study to definitively validate our automated metric findings and gain richer qualitative insights into user preferences. Furthermore, a direct quantitative comparison against a robust reconstruction-based method like DreamBooth+LoRA, using the same evaluation metrics, will provide a clearer benchmark for the gains achieved through preference-based alignment. Statistical significance of these comparisons will be rigorously assessed using Wilcoxon signed-rank tests.

Looking further ahead, with more resources, we would expand the evaluation to diverse artistic styles beyond the Laion Art subset and test our methods on different base diffusion models to assess generalizability. Incorporating multi-attribute and multimodal preference datasets, such as VisionReward or RLAIF-V as discussed in Section 4, could allow for more disentangled stylistic control and richer supervision. We also plan to explore more advanced human utility functions and sophisticated self-play mechanisms. Finally, a more exhaustive hyperparameter search and investigation into the impact of LoRA rank and training duration could yield further improvements, paving the way for highly controllable and personalized generative models for critical downstream applications in art and design.

7. Appendices

7.1. Text Prompts

No.	Prompt
1.	A photo of a cat with blue eyes
2.	A small cottage in the countryside
3.	A glass of water on a wooden table
4.	Portrait of a woman with flowers in her hair
5.	A futuristic city skyline at sunset

7.2. Stylized Prompts

No.	Prompt
1.	A watercolor painting of a cat with blue eyes, artistic, dreamy, soft brushstrokes
2.	An oil painting of a cozy cottage in impressionist style, vibrant colors, thick impasto
3.	A still life oil painting of a glass of water, Dutch Golden Age style, dramatic lighting
4.	A Renaissance portrait of a woman with flowers in her hair, ornate details, sfumato technique
5.	A cyberpunk digital art of a city skyline at sunset, neon colors, volumetric lighting

8. Contributions & Acknowledgments

All team members contributed significantly to the conceptualization, implementation, and analysis of this project. The primary responsibilities were distributed as follows:

- Wendy Yin led the implementation and training process of the core model. She also helped establish the connection between modern alignment techniques and economic utility theory. She helped develop the baseline models, implemented the loss functions for the Diffusion-KTO & SPIN-Diffusion alignment variants and conducted the hyperparameter tuning experiments.

- Yicheng Zhang led the initial literature and research review, focusing on the literature on preference alignment. He also designed the overall experimental framework and formulated the hypotheses of our project. Yicheng implemented the BCE LoRA alignment variant and developed the scripts for generating images from the final trained models.

- Yiwen Zhang led the development of the data and evaluation pipelines. He was responsible for curating the Laion Art dataset and implementing the logic to process the dataset's aesthetic scores into binary preference labels for training. Yiwen implemented the Diffusion-DPO alignment variant and the evaluation framework for calculating quantitative scores.

All team members contributed to writing and proofreading of the final report.

References

[1] S. Afriat. The construction of a utility function from demand data. *International Economic Review*, 8(1):67–77, 1967.

[2] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Kto: Model alignment as prospect theoretic optimization. 2024.

[3] fantasyfish. Laion-art. `https://huggingface.co/datasets/fantasyfish/laion-art`, 2023.

[4] E. Hu, Y. Shen, P. Wallis, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

[6] S. Li, K. Kallidromitis, A. Gokul, Y. Kato, and K. Kozuka. Aligning diffusion models by optimizing human utility. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024.

[7] D. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[9] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8821–8831, 2021.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[12] N. Ruiz, Y. Li, V. Jampani, et al. Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] C. Schuhmann, G. Couairon, R. Beaumont, R. Vencu, and LAION Community. Laion-aesthetics: Large-scale clip-based aesthetic scores. `https://laion.ai/blog/laion-aesthetics/`, 2022.

[14] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[15] D. Vera Nieto, L. Celona, and C. Fernandez-Labrador. Understanding aesthetics with language: A photo critique dataset for aesthetic assessment. *arXiv preprint arXiv:2206.08614*, 2022.

[16] E. Wallace, B. Bui, T. Varis, A. Kirubarajan, R. Bommasani, et al. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[17] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[18] J. Xu, Y. Huang, J. Cheng, Y. Yang, J. Xu, Y. Wang, W. Duan, S. Yang, Q. Jin, S. Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[19] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

[20] H. Yang, Y. Gong, and Y. Wang. D3po: Efficient preference alignment of diffusion models without reward networks. *arXiv preprint arXiv:2312.01234*, 2023.

[21] W. Yang, Q. Liu, and S. Li. Fifa: Filtering human feedback data for faster preference alignment. *arXiv preprint arXiv:2410.10166*, 2024.

[22] X. Yuan and H. Zhang. Self-play fine-tuning of diffusion models for text-to-image alignment. *arXiv preprint arXiv:2402.10210*, 2024.