Extended Abstract

Motivation Most text-to-image diffusion models like Stable Diffusion and DALL-E are good at generating high-fidelity images but often struggle to adhere to specific or niche artistic styles. Such limitation is mainly due to the fact that their pre-training data tend to average over stylistic nuances, resulting in images that may be aesthetically pleasing but misaligned with an artist's specific requirement. Addressing this style-specific alignment gap is crucial for downstream applications such as concept art, personalized game assets, and cultural heritage preservation, where detailed artistic fidelity is essential. This project investigates whether reinforcement learning methods combined with human feedback objectives can align these models with fine-grained artistic preferences.

Method The core approach of the project involves fine-tuning a Stable Diffusion v1.5 checkpoint using Low-Rank Adaptation (LoRA), a parameter-efficient technique that updates only a small fraction of the model's parameters. We compare several preference-alignment strategies, including a baseline using a simple binary cross-entropy loss (BCE LoRA), Direct Preference Optimization (DPO), Kahneman-Tversky Optimization (KTO), and Self-Play fIne-tuNing (`SPIN`). All methods were implemented within a unified LoRA framework to isolate the impact of the optimization objective.

Implementation We use the Laion Art subset as the primary training data, which aligns with the project's focus on niche artistic styles. To train the preference-based models, the dataset's existing aesthetic scalar scores were converted into deterministic win/loss labels, creating binary feedback without the need for manual annotation.

All fine-tuning started from the same Stable Diffusion v1.5 checkpoint and only updated rank-8 LoRA adapters in the U-Net. The training utilized an AdamW optimizer with a constant learning rate of 1e-4 for LoRA weights. All models were trained for 10,000 steps, which included a warm-up phase of 500 steps to ensure initial training stability. Due to computational limitations, an exhaustive hyperparameter search was not conducted. The experiments were run on a single A100-80GB GPU, with mini-batch sizes ranging from 4 to 8 samples to maximize utilization. To assess model performance, we utilized two primary automated metrics: CLIP Score and PickScore.

Results Quantitative evaluation using automated metrics showed that advanced alignment methods significantly outperformed baselines. SPIN-Diffusion achieved the highest human preference score (PickScore) with a mean of 21.95, closely followed by Diffusion-KTO at 21.80. Both substantially surpassed the original Stable Diffusion v1.5 (21.25) and the BCE LoRA baseline (20.74). All models maintained comparable CLIP Scores, indicating that the improvements in stylistic alignment did not compromise semantic integrity. Qualitatively, the Diffusion-KTO model produced images with finer details and more vivid colors compared to the base model.

Discussion The results strongly support the hypothesis that human utility optimization is more effective than naive preference learning. The success of SPIN-Diffusion demonstrates that a self-play strategy, which generates its own strong negative pairs, is highly effective in refining alignment without new human data. Furthermore, Diffusion-KTO exceeded the performance of Diffusion-DPO while requiring only simpler binary feedback (likes/dislikes) instead of pairwise comparisons, highlighting its data efficiency. The poor performance of the BCE LoRA model suggests that the choice of optimization objective is critical for successful alignment.

Conclusion The project demonstrates that aligning diffusion models using direct human utility optimization is a promising and efficient path to achieving high-fidelity stylistic control. Methods like SPIN-Diffusion and Diffusion-KTO, combined with parameter-efficient fine-tuning, significantly improve adherence to nuanced artistic styles over standard models and simpler baselines. These findings are critical for enabling downstream applications in art and design that demand specific aesthetic control.

Aligning Text-to-Image Diffusion Models using Reinforcement Learning from Human Utility

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

Yiwen Zhang
Department of Computer Science
Stanford University
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 applying reinforcement learning (RL) principles through methods that optimize expected human utility based on feedback 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 (Rombach et al., 2022) and DALL-E (Ramesh et al., 2021) 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. However, despite impressive breadth, these models remain coarse instruments when users demand adherence to *highly specific* or *niche* artistic styles. Constrained by the heterogeneous signal—and often the dominant one—in their pretraining data, they tend to average over stylistic nuances, producing 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, and reinforcement learning from human utility. We mainly study whether *parameter-efficient* diffusion model can be updated using Low-Rank Adaptation (LoRA) adapters to align to niche, fine-grained artistic preferences *using human-feedback objectives alone*, and investigate whether such alignment translates into measurable gains over both reconstruction-only finetuning 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 compact 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 $\mathcal{M}_0$, and (b) state-of-the-art preference-alignment methods like Diffusion-DPO, Diffusion-KTO, and SPIN-Diffusion. 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 compared to methods requiring explicit pairwise comparisons.
- **Advanced Optimization Objectives:** We systematically compare Diffusion-KTO & SPIN-Diffusion 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**, 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 in three interconnected domains: the creation of preference datasets, the development of optimization objectives to take advantage of these 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 crowdsourced pairwise votes, allowing systematic comparison of model outputs (Kirstain et al., 2023). ImageRewardDB extended this idea, collecting 137k expert comparisons and distilling them into a CLIP-based reward model (Xu et al., 2023). Human Preference Score v2 (HPSv2) was further scaled to 800 k comparisons and established a robust automatic metric (Wu et al., 2023).
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 image and video generation (Xu et al., 2024).

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 (Xu et al., 2023). Direct Preference Optimization (DPO) (Rafailov et al., 2024), 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 (Wallace et al., 2024). D3PO further reduces memory overhead by operating in the denoising latent

space (Yang et al., 2023). Our work heavily leverages a successor to these methods, Diffusion-KTO. Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) 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 can maximize expected human utility using only binary feedback (e.g., likes/dislikes), which dramatically simplifies the data collection process. (Li et al., 2024).

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 (Yuan and Zhang, 2024). Moreover, FiFA proposes automated filtering that can accelerate DPO training by two orders of magnitude, making the alignment process faster and more efficient (Yang et al., 2024).

Connection to historical development of utility function: Under the von-Neumann–Morgenstern axioms, binary feedback or pairwise comparisons can be embedded 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 (McFadden, 1974). Moreover, Afriat's revealed-preference theorem guarantees that, absent preference cycles, a continuous, monotone utility rationalizes any finite set of binary choices (Afriat, 1967); 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 Method

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 three 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 $2dr$ (Hu et al., 2021). 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 (Song et al., 2021), 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 (Xu et al., 2023), 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) 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 (Rafailov et al., 2024). 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

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

where r_θ is the per-image implicit reward function parameterized by the diffusion model itself (Wallace et al., 2024). 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.

(ii) 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:

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

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:

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

In the estimator, λ serves as a baseline to reduce the variance of the gradient estimates, leading to more stable training (Li et al., 2024). KTO's ability to learn from simple, unpaired human feedback makes it a highly data-efficient and more flexible compared with other alignment methods.

(iii) 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 (Yuan and Zhang, 2024).

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 Binary-Preference LoRA baseline, (H2) that KTO matches DPO despite needing only unpaired likes, and (H3) that SPIN yields further gains by self-generating hard negatives.

4 Experimental Setup

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).

4.1 Dataset

For this project we adopt the **Laion Art subset** as our core training source (fantasyfish, 2023). 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.

4.2 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.

4.3 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) (Radford et al., 2021). 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 (Kirstain et al., 2023) 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 Results

5.1 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.

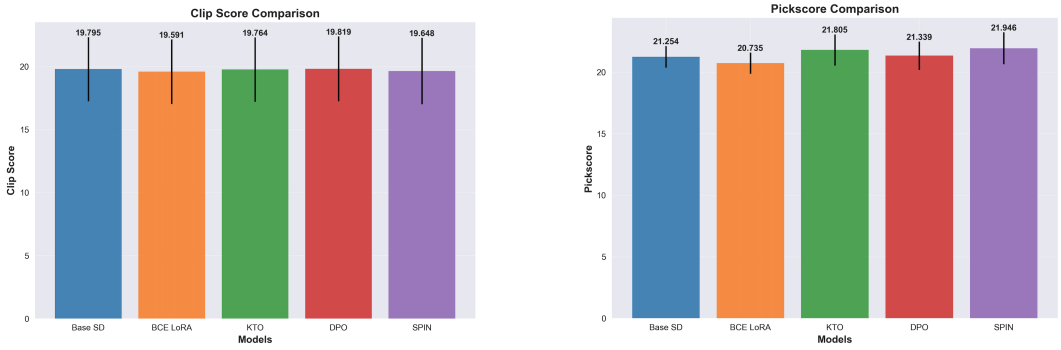
The models were evaluated on a diverse set of prompts, exemplified in Appendix A (general photographic) and Appendix B (stylized counterparts). This set was designed to cover varied subjects and styles, allowing for assessment of both general artistic rendering and adherence to specific stylistic keywords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

5.1.1 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

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

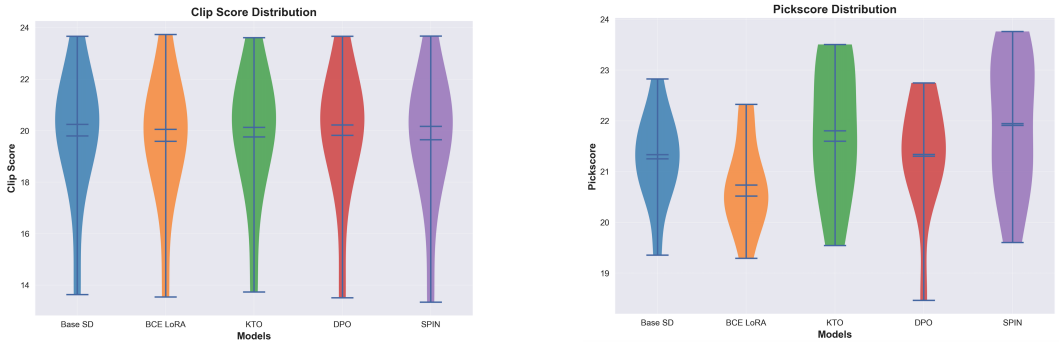
(b) PickScore Comparison

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

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.

5.1.2 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.



(a) CLIP Score Distribution

(b) PickScore Distribution

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

5.2 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.

(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.

5.3 In-Depth Analysis

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 3.3:

DPO and KTO outperform the Binary-Preference LoRA 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.

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.

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.

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.

6 Discussion

The success of KTO is particularly promising for further research 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.

7 Conclusion

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. The statistical significance of these comparisons will be rigorously assessed using Wilcoxon signed-rank tests.

8 Team Contributions

- **Wendy Yin:** led the implementation and training process of the core model; helped establish the connection between modern alignment techniques and economic utility theory; formulated the hypotheses of our project; developed the baseline models, implemented the loss functions for the Diffusion-KTO & SPIN-Diffusion alignment variants and conducted the hyperparameter tuning experiments.
- **Yiwen Zhang:** led the initial literature and research review; developed the data and evaluation pipelines; designed the overall experimental framework; developed the scripts for

generating images from the final trained models, implemented the BCE LoRA & Diffusion-DPO alignment variants and the evaluation framework for calculating quantitative scores.

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

Changes from Proposal We researched one additional alignment variant SPIN-Diffusion that is not mentioned in the proposal, and compared its performance with Diffusion-DPO & Diffusion-KTO to investigate the benefits of self-generating hard negatives.

References

Sidney N. Afriat. 1967. The Construction of a Utility Function from Demand Data. *International Economic Review* 8, 1 (1967), 67–77.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. arXiv:2402.01306 [cs.LG] `https://arxiv.org/abs/2402.01306`

fantasyfish. 2023. *LAION-Art*.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. *arXiv preprint arXiv:2305.01569* (2023). `https://arxiv.org/abs/2305.01569`

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. 2024. Aligning Diffusion Models by Optimizing Human Utility. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*. `https://arxiv.org/abs/2404.04465`

DanielL. McFadden. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, Paul Zarembka (Ed.). Academic Press, 105–142.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] `https://arxiv.org/abs/2103.00020`

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] `https://arxiv.org/abs/2305.18290`

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 8821–8831.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*.

Eric Wallace, Brian Bui, Tony Varis, Anand Kirubarajan, Rishi Bommasani, et al. 2024. Diffusion Model Alignment Using Direct Preference Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. `https://arxiv.org/abs/2311.12908`

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341* (2023). `https://arxiv.org/abs/2306.09341`

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. 2024. VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation. *arXiv preprint arXiv:2412.21059* (2024). https://arxiv.org/abs/2412.21059

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. https://arxiv.org/abs/2304.05977

Huan Yang, Yutong Gong, and Yuchen Wang. 2023. D3PO: Efficient Preference Alignment of Diffusion Models without Reward Networks. *arXiv preprint arXiv:2312.01234* (2023). https://arxiv.org/abs/2312.01234

Wenjun Yang, Qilin Liu, and Shuo Li. 2024. FiFA: Filtering Human Feedback Data for Faster Preference Alignment. *arXiv preprint arXiv:2410.10166* (2024). https://arxiv.org/abs/2410.10166

Xin Yuan and Hao Zhang. 2024. Self-Play Fine-Tuning of Diffusion Models for Text-to-Image Alignment. *arXiv preprint arXiv:2402.10210* (2024). https://arxiv.org/abs/2402.10210

A 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

B 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