# Aligning Diffusion Models with Online Preference Learning

Tianyu Zhan
New York University

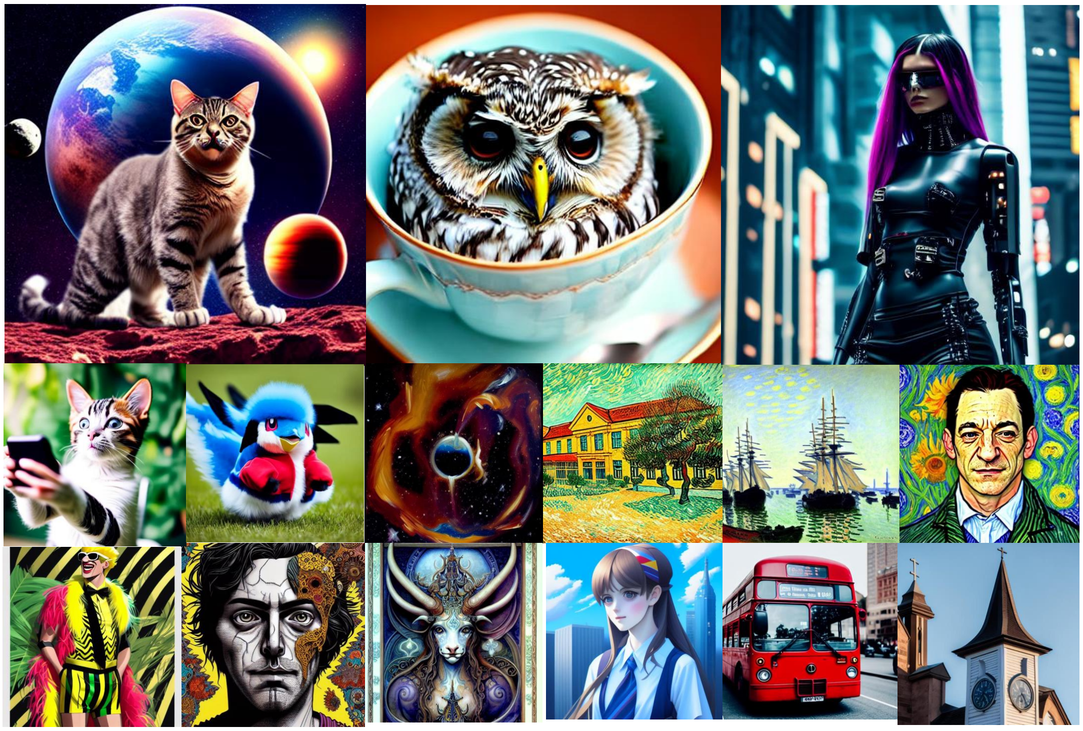

Figure 1. **Overview of EXPO.** Our proposed EXPO framework significantly improves text-to-image generation quality through iterative online training. By combining limited offline data with self-generated samples via explicit exploration mechanisms, EXPO achieves superior performance compared to models trained solely on offline data. The samples above demonstrate EXPO's enhanced generation capabilities on unseen prompts across diverse styles and content.

## Abstract

*Reinforcement learning from human feedback (RLHF) has shown remarkable success in large language models (LLMs), yet its application to diffusion models remains relatively underexplored. Direct Preference Optimization (DPO)[19] provides a more stable alternative to policy-based methods like Proximal Policy Optimization (PPO)[22], but recent works such as DiffusionDPO [26] still exhibit limited improvements when relying solely on offline data.To overcome this challenge, we propose* ***EXPO****, a hybrid DPO training framework for diffusion models that integrates acitve e****X****ploration into preference* ***O****ptimization. EXPO enables iterative RLHF training by alternating between exploiting historical data and exploring uncertain regions of the prompt space with preferece-based exploartion. Experiments on high-fidelity open-source datasets, including Pick-a-pic v2 and HPDv2, using Stable Diffusion v1-5 (SD1.5) demonstrate that EXPO significantly significantly outperforms purely offline baselines. Moreover, our exploration mechanism shows promise in mitigating social biases through selective data acquisition, underscoring the broad applicability of EXPO in diffusion-based generative modeling.*

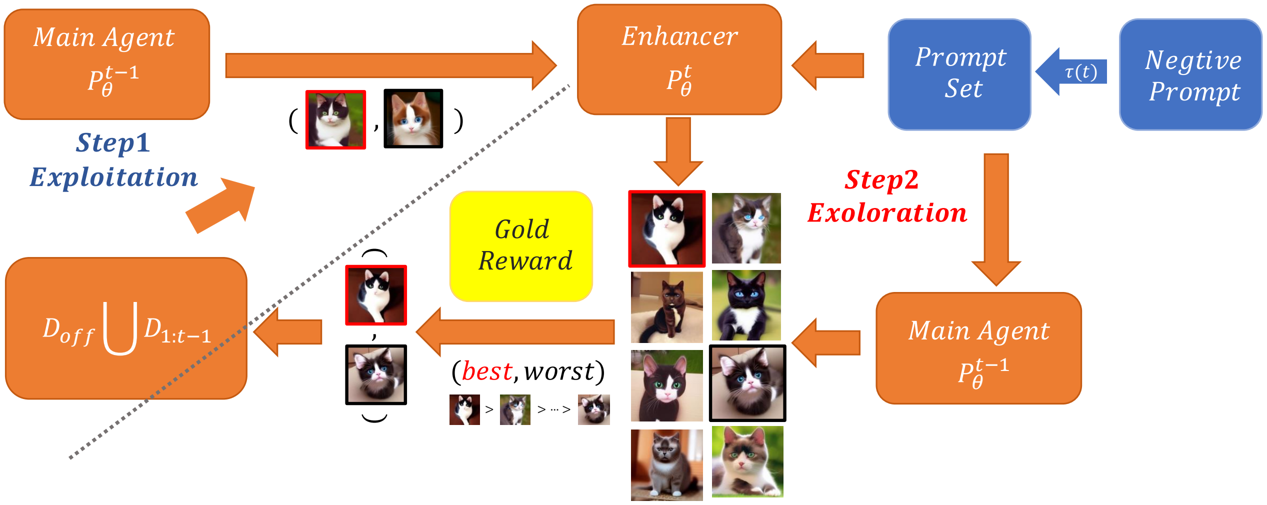

Figure 2. **Overview of our EXPO framework.** Our EXPO framework demonstrates superior image generation quality compared to Stable Diffusion v1.5 and DiffusionDPO across diverse prompts. EXPO achieves these improvements while using only 15% of the training data and 1/4 of the training steps, highlighting its efficiency and effectiveness. The samples above showcase enhanced alignment with text prompts and improved aesthetic quality.

## 1. Introduction

Diffusion models have emerged as a powerful class of generative models, demonstrating remarkable capabilities in creating rich and aesthetically pleasing images [14, 23, 24]. Trained on large-scale offline datasets, these models exhibit strong generalization abilities, making them valuable foundational models for various downstream tasks. However, despite their success, diffusion models often produce outputs that misalign with human prompts or lack certain desired aesthetic qualities.

To address these issues, several approaches have been explored. One line of work focuses on designing new architectures [10, 18], which typically entails retraining models from scratch using larger and higher-quality datasets [12, 21]. While these methods can enhance performance, they are resource-intensive, requiring substantial computational resources and time, making them less practical for widespread use.

Alternatively, fine-tuning pre-trained diffusion models has been proposed as a more efficient solution. Methods such as supervised fine-tuning on high-quality images [13, 20] aim to improve alignment with human prompts. However, these approaches are constrained by the quality and diversity of the offline data, limiting their effectiveness. Some works attempt to enhance models by generating additional data through self-sampling, achieving better performance but still facing limitations inherent in the offline training paradigm.

In the realm of language models, Reinforcement Learning from Human Feedback (RLHF) has become a standard approach for aligning models with human preferences, effectively bridging the gap between model outputs and user expectations [2, 3, 17]. RLHF methods, such as Proximal Policy Optimization (PPO)[22], have been instrumental in training large language models like LLaMA[9, 25]. Direct Preference Optimization (DPO) [19], a variant of RLHF, offers improved training stability and has shown promise in this context.

Inspired by the success of RLHF in language models, researchers have begun to explore its application to text-to-image diffusion models. Initial attempts, such as RAFT [8], employed supervised fine-tuning and rejection sampling to align model outputs with human preferences. However, these methods can be inefficient and may degrade image quality due to overfitting or lack of diversity. RL-based approaches like DDPO and DPOK [4, 11] formulate the alignment problem within a reinforcement learning framework using policy gradients. While these methods represent significant progress, they often require extensive computational resources and face challenges scaling to larger datasets due to over-optimization issues.

Diffusion-DPO [26] proposes optimizing diffusion models using DPO objectives derived from the Evidence Lower Bound (ELBO), enabling more efficient training. Nonetheless, their study is limited to offline settings, which may not fully capture the benefits of online data collection and iterative learning. Other works, such as PRDP [7], integrate reward models with PPO objectives for online training but suffer from challenges like reward hacking and increased memory overhead. IterComp [31] explores compositional learning by ensembling multiple state-of-the-art diffusion models but does not directly address alignment with human preferences.

In this work, we introduce **EXPO**, a novel exploratory iterative Direct Preference Optimization method for diffusion models that incorporates explicit exploration mechanisms through negative prompts. Our approach enables iterative RLHF training, effectively leveraging both offline and online data collection to improve model alignment with human preferences. By balancing exploitation and exploration, EXPO addresses the limitations of previous methods, achieving superior performance in large-scale settings.

Our main contributions are as follows:

1. We propose **EXPO**, a novel training framework that incorporates asymmetric explicit exploration mechanisms into iterative DPO training for diffusion models, enabling an effective balance between exploitation and exploration during training.
2. We demonstrate through extensive experiments that EXPO significantly outperforms existing offline methods on large-scale, high-fidelity datasets including Pick-a-Pic v2 [15] and HPDv2 [27]. Our method exhibits strong scalability and consistently improves performance on both in-distribution and out-of-distribution prompts across training iterations.
3. We show that our exploration strategy can effectively guide model outputs through curated data, providing a principled approach to mitigate social biases while maintaining model safety through online exploration mechanisms.

## 2. Related Work

**Reinforcement Learning for Language Models** Reinforcement Learning from Human Feedback (RLHF) has been crucial in aligning large language models with human preferences, leading to advancements in models like ChatGPT [1]. The predominant approach uses Proximal Policy Optimization (PPO) [22], but it requires extensive resources and careful tuning [6], limiting its accessibility.

To address this, alternative methods like Rejection Sampling Fine-Tuning (RAFT) [8] and direct preference learning algorithms such as DPO [19] have been developed. These methods optimize directly on preference data, bypassing explicit reward models, and have significantly advanced RLHF, especially in open-source models.

Recent works highlight the effectiveness of **on-policy sampling** and **online exploration** in enhancing preference learning [28, 29]. By balancing exploration and exploitation, these methods enable models to discover new strategies and improve alignment with human preferences. However, most algorithms focus on single-turn interactions, leaving multi-turn scenarios largely unexplored.

**Reinforcement Learning for Diffusion Models** Applying RLHF to diffusion models presents unique challenges due to the sequential nature of the denoising process. Unlike language models that produce single-turn outputs, diffusion models involve multi-step generation, making direct application of RLHF techniques impractical. Some methods [16] use reward-weighted likelihood maximization but overlook the sequential aspects.

To address this, approaches like DDPO [4] and DPOK [11] model the denoising process as a Markov Decision Process and apply policy gradient algorithms. Others leverage differentiable reward models [30] to optimize via backpropagation. However, these methods often rely on offline datasets, require differentiable rewards, or suffer from increased memory overhead and over-optimization.

Our proposed method, **EXPO**, introduces a stable and efficient online training algorithm that balances exploration and exploitation without relying on differentiable rewards. It addresses the unique challenges of aligning diffusion models with human preferences, marking the first exploration of this balance in diffusion model alignment.

## 3. Preliminaries

In this section, we will introduce the basic concepts of diffusion models and the related RLHF methods.

**Diffusion Models** Diffusion Models provide a powerful framework for modeling complex data distributions $q(x_0)$ through a Markov process. The key idea is to gradually transform data points through a sequence of denoising steps, where each step $x_t$ is derived from its predecessor $x_{t-1}$. And performing the below objective to optimize the models:

$$\mathbb{E}_{x_{0,t,\epsilon,t}}, c\left[\omega_t \|\epsilon - \epsilon_\theta(x_t, c, t)\|^2\right] \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,1)$, $x_t \sim q(x_t|x_0) = \mathcal{N}(x_0; \alpha_t x_0, \sigma_t^2 I)$, $\omega_t$ is a predefined weighting function.

**DPO for Diffusion Models** The Bradley-Terry (B-T) model provides a principled framework for modeling human preferences with the implicit reward function $r^*(x, c)$. Given a pair of preference data $x_0$ and $x_0'$, the B-T model defines the probability that image $x_0$ is preferred over image $x_0'$ given condition $c$ as:

$$P(x_0 \succ x_0' \mid c) = \sigma\Big(r^*(x_0, c) - r^*(x_0', c)\Big) \tag{2}$$

We define the standard error as

$$SE(x, c) = \|\epsilon - \epsilon_\theta(x_t, t \mid c)\|_2^2.$$

Diffusion-DPO [26] proves that optimizing the RLHF objective is equivalent to minimizing the following objec-

**Algorithm 1** EXPO-II Training Pipeline

```
Require: Reward Model $r^*$, Offline Dataset $\mathcal{D}_{\text{off}}$, Pretrained Model $p_\theta$, Prompt Set $\mathcal{D}_\mathcal{C}$, Negative Prompt Set $\mathcal{D}_\mathcal{N}$, Images per Prompt $B$, Total Training Iterations $T$, Number of Prompts per Iteration $N$
Initialize $p_{\text{ref}} \leftarrow p_0 \leftarrow p_\theta$
for $t = 1$ to $T$ do
    $\mathcal{D}_{\text{hybrid}} \leftarrow \mathcal{D}_{\text{off}} \cup \mathcal{D}_{1:t-1}$
    Policy Update:                                  ▷ Exploit historical data
    Update policy parameters by calling the policy optimization oracle Eq. (4)
    if $t < T$ then                                 ▷ Explore new regions
        for each prompt $\mathbf{c}^n \sim \mathcal{D}_\mathcal{C}$ do
            Sample negative prompt $\mathbf{c}_{\text{neg}} \sim \rho(t, \mathcal{D}_\mathcal{N})$
            Augmented prompt: $\mathbf{c} \leftarrow \mathbf{c}^n + \mathbf{c}_{\text{neg}}$
            $\mathcal{D}_{\text{main}} \leftarrow \{(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_B) \sim p^{\text{main}}(\cdot|\mathbf{c})\}$
            $\mathcal{D}_{\text{ex}} \leftarrow \{(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_B) \sim p^{\text{ex}}(\cdot|\mathbf{c})\}$
            Collect data $\mathcal{D}_t \leftarrow \mathcal{D}_{\text{main}} \cup \mathcal{D}_{\text{ex}}$
        end for
    end if
end for
```

tive:

$$L_{\text{Diffusion-DPO}}(\theta) = -\mathbb{E}_{\substack{(x_0^w, x_0^t) \sim \mathcal{D}, t \sim U(0,T) \\ x_t^w \sim q(x_t^w | x_0^w), x_t^t \sim q(x_t^t | x_0^t)}} \log \sigma \Bigg( - \beta \Big( SE_\theta(x_t^w, t) - SE_{\text{ref}}(x_t^w, t) - \Big( SE_\theta(x_t^l, t) - SE_{\text{ref}}(x_t^l, t) \Big) \Big) \Bigg) \tag{3}$$

where $\text{SE}_\theta$ represents the current model and $\text{SE}_{\text{ref}}$ represents the reference model. We merge $T$ into $\beta$ and set the weighting function to be constant here.

## 4. Methodology

**Online Exploration for Preference Learning** We define a policy optimization oracle analogous to Xiong *et al.* [28]:

$$p^* = \arg\max_p \mathbb{E}_{c \sim \mathcal{D}_c, p} \Big[ r(\mathbf{x}_{0:T}) - \beta \, \text{KL} \left( p_\theta(\mathbf{x}_{0:T} | \mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_{0:T} | \mathbf{c}) \right) \Big], \tag{4}$$

As discussed in [28] and [5], the offline dataset cannot cover all possible regions in the policy space, leading to the model's inability to perform well on out-of-distribution data.

To address this, we define a main policy $p^{\text{main}}$, an exploration policy $p^{\text{ex}}$ and an uncertainty quantifier $\Gamma_t(x_0, c)$. We aim for the main policy to exploit existing information while the exploration policy explores regions with high uncertainty. We incorporate the uncertainty quantifier into Eq. (3) to form our new objective function:

$$L_{\text{EXPO}}(\theta) = -\mathbb{E}_{\substack{(x_0^w, x_0^l) \sim \mathcal{D} t \sim U(0,T) \\ x_t^w \sim q(x_t^w | x_0^w) x_t^l \sim q(x_t^l | x_0^l)}} \log \sigma \Bigg( -\beta T \Big( SE_\theta(x_t^w, t) - SE_{\text{ref}}(x_t^w, t) - \Big( SE_\theta(x_t^l, t) - SE_{\text{ref}}(x_t^l, t) \Big) \Big) \Bigg) + \Big( \Gamma_t(x_0^w, c) - \Gamma_t(x_0^l, c) \Big) \tag{5}$$

Defining the uncertainty quantifier $\Gamma_t(x_0, c)$ is the key challenge of our method. following [28], we adopt a heuristic approach to define our iterative exploration algorithm. We have also defined a explicit traning objective for exploitation later.

**Algorithm Design** We propose Algorithm 1, which outlines our training pipeline combining exploitation and exploration.

In each iteration, we perform the following steps:

1. **Exploitation**: We update the policy using both the offline dataset and the data collected in previous iterations. This step leverages historical information to improve the policy.
2. **Exploration**: We generate new data by sampling from the exploration policy. We augment prompts with negative samples to encourage the model to explore less certain regions, potentially discovering more informative data.
3. **Data Collection**: We collect the generated images and add them to the dataset for future iterations.

By alternating between exploitation and exploration, our algorithm efficiently improves the policy's performance on out-of-distribution data.

**Preference-based Exploration** From the noise-aware preference perspective, the Diffusion-DPO loss Eq. (3) aims to minimize $SE_\theta(x^w, c)$ and maximize $SE_\theta(x^l, c)$ respectively. The asymmetry of the $\log \sigma$ function allows $\beta$ to control the penalty for deviating from the reference distribution. A high $\beta$ results in a highly asymmetric distribution, disproportionately penalizing low $SE_\theta$ (for weak cases) and high $SE_\theta$ (for strong cases), thereby encouraging $p_\theta$ to make fewer mistakes. We can further define the noise-aware reward difference as:

$$\begin{aligned} r(c, x_0^A) - r(c, x_0^B) &= s(c, x_t^A) - s(c, x_t^B) \\ &= \beta \Big[ \big( SE_\theta^A - SE_{\text{ref}}^A \big) - \big( SE_\theta^B - SE_{\text{ref}}^B \big) \Big], \quad \forall c, t \end{aligned} \tag{6}$$

The training objective in Eq. (3) employs a $-\log\sigma$ function whose gradient approaches zero when the reward difference is large, which can lead to gradient saturation problems. In other words, if the reward difference is huge, the training signal will be extremely low.

Recent work in LLMs has shown promising directions with active exploitation mechanisms—many of which are based on augmenting the DPO loss with a reward-based optimistic bonus to encourage exploration. Inspired by this, we propose our preference-incentive optimization objective:

$$
\begin{aligned}
L_{\text{EXPO}}(\theta) = -\mathbb{E}_{\substack{(x_0^w, x_0^l)\sim\mathcal{D}, t\sim U(0,T) \\ x_t^w\sim q(x_t^w|x_0^w), x_t^l\sim q(x_t^l|x_0^l)}} \\
\log\sigma\left(-\beta T\left(s(c, x_t^w) - s(c, x_t^l)\right)\right) \\
+\ \alpha\,\mathbb{E}_{x_0\sim\mathcal{D}_{\text{ex}}, x_0'\sim\mathcal{D}_{\text{main}}}\left[\sigma\left(\beta(s(c, x_0) - s(c, x_0'))\right)\right].
\end{aligned}
\tag{7}
$$

where $\alpha$ is the exploration temperature and the exploration term aims to encourage the exploration policy to find more favorable results compared to the reference policy.

# 5. Experiments

## 5.1. Experimental Setup

We conduct comprehensive experiments to evaluate the effectiveness of our proposed method. Our evaluation is based on two large-scale human preference datasets: Pick-a-Pic v2 and HPDv2. The Pick-a-Pic v2 dataset comprises approximately 1M preference pairs crowdsourced from web users, while HPDv2 contains 600K image-text preference pairs curated through rigorous human evaluation by professional annotators. For automated evaluation metrics, we leverage both PickScore and HPS v2.1 as our reward model and evaluation metrics. To assess generalization capability, we utilize the HPS benchmark and PickScore test set. All experiments are implemented using the Stable Diffusion v1.5 architecture as our base model. We define four variants of our online training algorithm: EXPO-I, EXPO-II and EXPO-II-M. EXPO-II is the asymmetric training algorithm that we have introduced before. P means we use the PickScore as our reward model and trained on the Pick-a-Pic v2 dataset. H means we use the HPS v2.1 as our reward model and trained on the HPDv2 dataset. EXPO-II-M, we use the multi-reward model to generate data. As for EXPO-I, we follow the desing of RAFT[8], we trained the model iteratively, and the exploration and main agents remain ths same. For all of these variants, we select apply the negtive reporst with linear decay and select the best and worst of the generated images to update the model.

**Hyperparameters** For all of our experiments, we use a local batch size of 1 and tried different gradient accumulation steps between 16 and 128. Training on 8 Nvidia H100 GPUs, Thus our effective batch size is between 16 and 512. And for each iteration, we train the model for 2000 steps, which is about 10 hours on 4 Nvidia H100 GPUs. the total training time for each iteration is about 10 hours. So after 4 iterations, our effective training step is equal to the DiffusionDPO[26]. For all our experiments, we use the learning rate of 1e-8 and the weight decay of 1e-2 and beta (0.9, 0.999), adam episilon is 1e-8 with AdamW optimizer. All of our experiments are trained with 25% warmup steps.

For sampline, all of our experiments are conducted using DDPM[14] sampler with 50 inference steps, and a classifier-free guidance scale of 7.5. We keep the $\beta = 5000$ for all of our experiments as we we want to increase the cost of awat from the reference model.

**Evaluation Metrics** We use the 500 unique prompts from Pick-a-pic v2 validation set [15] to automatically evaluate the performance of our method during training.

And we pick-a-picv2 test set, which contains 500 unique prompts, and the HPS benchmark[27], which have 4 categories: anime, concept art, photo, and paintings and 3200 text-image pairs, to evaluate the performance of our method. We use the reward model PickScore and HPSv2.1 to evaluate the quantiative performance.

We then conduct the hiring human lablers to evalue our method with respect to three aspects: 1. The aesthetic of the generated images. 2. The alignment of the generated images with the prompt. 3. General quality of the generated images.

**Implementation Details** We begin by aggregating all prompts from the Pick-a-Pic v2 and HPDv2 training sets. We then eliminate duplicate prompts and those exceeding 77 CLIP tokens, resulting in 103,684 prompts from HPDv2 and 56,899 prompts from Pick-a-Pic v2.

Initially, we randomly sample 81,920 offline text-image pairs from each dataset to form our initial offline dataset. For each iteration, we randomly sample 20,480 prompts and generate images with a batch size of 4 for each agent. The model is trained for a total of 5 iterations to accumulate an equivalent amount of online text-image pairs.

## 5.2. Quantitative Results

For offline method such as DiffusionDPO, DiffusionKTO and SPO, we use the the HPDv2 training set and Pick-a-Pic v2 test set, and train the model with approximately the same training steps as our model for a fair comparison. First, we evaluate the performance of our method on both the HPS benchmark[27] and Pick-a-pic v2[15] training sets. As shown in table 1. Our online iterative traninig methods

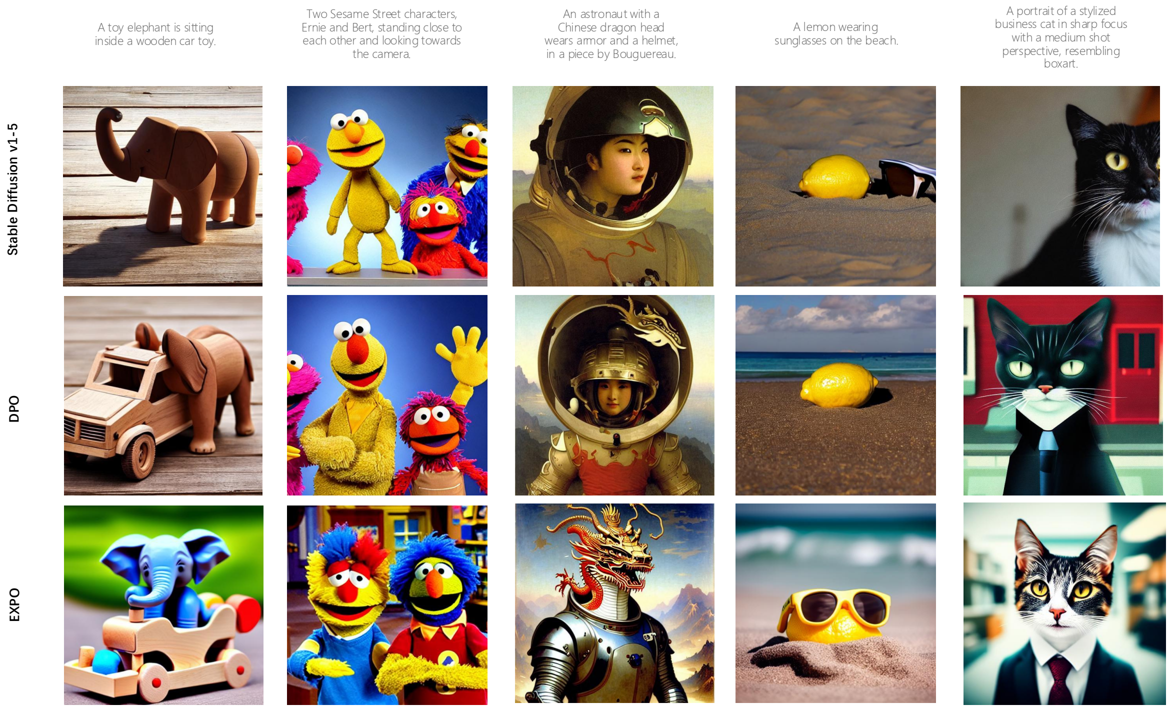
A toy elephant is sitting inside a wooden car toy.
Two Sesame Street characters, Ernie and Bert, standing close to each other and looking towards the camera.
An astronaut with a Chinese dragon head wears armor and a helmet, in a piece by Bouguereau.
A lemon wearing sunglasses on the beach.
A portrait of a stylized business cat in sharp focus with a medium shot perspective, resembling boxart.
Stable Diffusion v1-5
DPO
EXPO

Figure 3. **Qualitative comparison with baseline models.** Our EXPO framework demonstrates superior image generation quality compared to Stable Diffusion v1.5 and DiffusionDPO across diverse prompts. EXPO achieves these improvements while using only 15% of the training data and 1/4 of the training steps, highlighting its efficiency and effectiveness. The samples above showcase enhanced alignment with text prompts and improved aesthetic quality.

Table 1. **Performance comparison on HPS benchmark.** Our method significantly outperforms previous methods across all categories. Best results are shown in **bold**.

| **Method** | **HPS v2.1** | | | | | **PickScore** |
|---|---|---|---|---|---|---|
| | Animation | Concept-art | Paintings | Photos | Avg | Pick-a-Picv2 test |
| SD1.5 | 24.84 | 23.93 | 23.82 | 25.02 | 24.40 | 0.1473 |
| SFT | 25.73 | 24.82 | 24.68 | 25.21 | 25.11 | 0.1475 |
| DiffusionDPO-P | 23.89 | 22.03 | 22.18 | 24.82 | 23.23 | 0.1482 |
| DiffusionDPO-H | 26.88 | 25.98 | 25.76 | 25.67 | 26.07 | 0.1477 |
| DiffusionKTO | 29.91 | 29.44 | 29.29 | 28.36 | 29.25 | 0.1679 |
| SPO | 28.73 | 28.18 | 28.23 | 26.03 | 27.79 | 0.1711 |
| EPO-I-P | 29.13 | 28.17 | 28.21 | 27.02 | 28.13 | 0.1697 |
| EPO-I-H | 30.65 | 30.47 | 30.44 | 28.80 | 30.09 | 0.1682 |
| EPO-II-P | 29.53 | 28.60 | 28.36 | 26.77 | 28.31 | 0.1701 |
| EPO-II-H | 30.47 | 30.51 | 30.31 | 28.54 | 29.96 | 0.1695 |
| EPO-II-MP | 30.73 | 30.34 | 30.23 | 28.28 | 29.90 | **0.1726** |
| EPO-II-MH | **30.86** | **30.76** | **30.94** | **28.71** | **30.32** | 0.1714 |

are generally euqal or outperform the offline methods. Specially, our EXPO-II-MH achieves the best performance on HPS benchmark and EXPO-II-P achieves the best performance on Pick-a-Pic v2 test set. We noticed that our models are trained which dataset, the socre of another benchmark also increases. Which indicates that our model are truly explored the policy space, and showcasing the effectiveness of our method.

**Negtive Prompts Enhance and Accelerate Aligment** We also conduct ablation studies to analyze how negative prompts affect the performance of our method. As shown in Figure 4, we observe that the performance of our method is sensitive to negative prompts. Without using negative

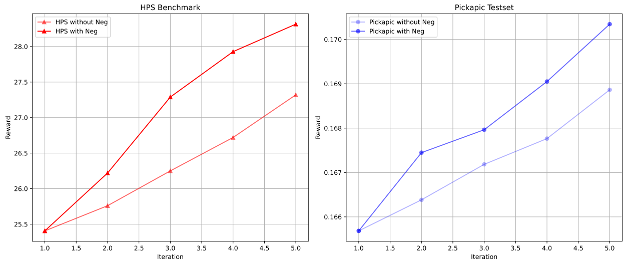

Figure 4. Ablation study on the effect of negative prompts. The results show significant improvements in the first four iterations when negative prompts are used.

prompts, the score of our method increases slowly. When we constrain a constant rate to add negative prompts to the model, the score is comparable with the method using $\tau(t)$ to manage negative prompt addition. However, we observe from Figure 4 that as training iterations increase, the constant rate approach quickly suffers from over-optimization problems, generating images with overly saturated colors.

**Negative Prompts Enhance Performance in Early Iterations** We find that incorporating negative prompts can significantly enhance the performance of the text-to-image model, especially in the early iterations. As shown in Figure 4, which uses the image from neg.pdf, the ablation study demonstrates that adding negative prompts results in noticeable improvements compared to not using them in the first four iterations. This indicates that negative prompts play a crucial role in accelerating the model's alignment and performance enhancement during the initial training phase.

**Model Performance consistently increase with iteration** We find that our method can consistently improve the performance of text-to-image model when we scale up the iteration. As shown in Figure 5, we show the trade-off between reward and KL divergence during training. The x-axis represents the KL divergence between the current policy and the reference policy, and the y-axis represents the reward score. We can observe that our method can achieve better reward with smaller KL divergence compared to offline methods. This indicates that our method can explore the policy space more efficiently and find better policies with less distribution shift.

**Qualitative Results** Here we show some qualitative results of our method and the compared methods. As we can see in figure, with iteration increases, the aesthetic perfromance of our model consistently become much better.

Also we showed how our model align with our prompts better than before.

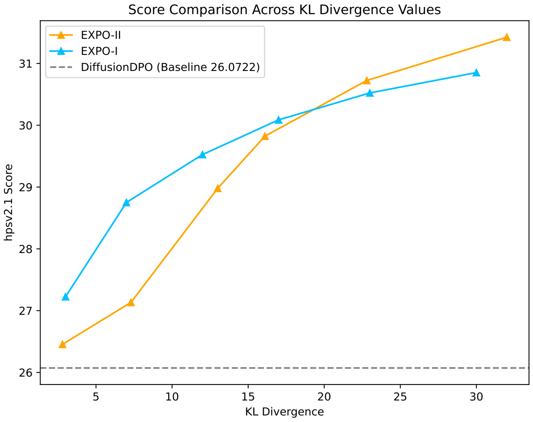

Figure 5. Reward-KL trade-off curve during training. Our method achieves better reward with smaller KL divergence compared to offline methods, indicating more efficient policy exploration.

As a most-relevent work PRDP[7], we don't have access to its code and pretraiend model, so we compare their model with some specific prompts that shows on their paper, we can see at apoendix B.

**Model safety** As most of Image generation models are pre-trained on web-scale data, it inevitablu include some social bias, and also, unfiltered NSFW information. We hire 10 professional workers to evaule our model safety capabilities compared with pre-trained sdv1.5 and DiffusionDPO. The results shows ...

**Human Evaluation** We hired 10 professional annotators to evaluate the generated images with respect to three aspects: 1. The aesthetic of the generated images. 2. The alignment of the generated images with the prompt. 3. General quality of the generated images. Here we show our winning rate compared to DiffusionDPO.

## References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Con-

stitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. 2

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2, 3

[5] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023. 4

[6] Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. On the weaknesses of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1907.01752*, 2019. 3

[7] Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7423–7433, 2024. 2, 7

[8] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. 2, 3, 5

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[10] 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. 2

[11] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[15] 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. 3, 5

[16] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[18] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[19] 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, 2024. 1, 2, 3

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1, 2, 3

[23] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[25] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[26] 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*, pages 8228–8238, 2024. 1, 2, 3, 5

[27] 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. *arXiv preprint arXiv:2306.09341*, 2023. 3, 5

[28] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024. 3, 4

[29] Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023. 3

[30] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[31] Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv preprint arXiv:2410.07171*, 2024. 3