HUMAN-FEEDBACK EFFICIENT REINFORCEMENT LEARNING FOR ONLINE DIFFUSION MODEL FINETUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable generation through Stable Diffusion (SD) fine-tuning aims to improve fidelity, safety, and alignment with human guidance. Existing reinforcement learning from human feedback methods usually rely on predefined heuristic reward functions or pretrained reward models built on large-scale datasets, limiting their applicability to scenarios where collecting such data is costly or difficult. To effectively and efficiently utilize human feedback, we develop a framework, HERO, which leverages online human feedback collected on the fly during model learning. Specifically, HERO features two key mechanisms: (1) *Feedback-Aligned Representation Learning*, an online training method that captures human feedback and provides informative learning signals for fine-tuning, and (2) *Feedback-Guided Image Generation*, which involve generating images from SD's refined initialization samples, enabling faster convergence towards the evaluator's intent. We demonstrate that HERO is $4\times$ more efficient in online feedback for body part anomaly correction compared to the best existing method. Additionally, experiments show that HERO can effectively handle tasks like reasoning, counting, personalization, and reducing NSFW content with only 0.5K online feedback.

1 INTRODUCTION

Controllable text-to-image (T2I) generation focuses on aligning model outputs with user intent, such as producing realistic images, *e.g.*, undistorted human bodies, or accurately reflecting the count, semantics, and attributes specified by users. To tackle this problem, a common paradigm involves fine-tuning latent diffusion models (DM) like Stable Diffusion (SD; Rombach et al., 2022) using supervised fine-tuning (SFT; Lee et al., 2023), which mostly learn from pre-collected, offline datasets. To further enhance the alignment, online reinforcement learning (RL) fine-tuning methods (Fan et al., 2024; Black et al., 2024) utilize online feedback that specifically evaluates the samples generated by the model during training. With such dynamic guidance provided on the fly, these methods demonstrate superior performance on various T2I tasks, such as aesthetic quality improvement. Yet, these approaches rely on either predefined heuristic reward functions or pretrained reward models learned from large-scale datasets, which could be challenging to obtain, especially for tasks involving personalized content generation (*e.g.*, capturing cultural nuances) or concepts like specific colors or compositions.

To address the above issue, Yang et al. (2024b) introduces D3PO, an alternative method that directly leverages online human feedback for fine-tuning diffusion models. Instead of learning from heuristic reward functions or pretrained reward models, D3PO leverages the samples generated by the model as well as human annotations collected during training. With online human feedback, D3PO addresses various tasks, such as distorted human body correction and NSFW content prevention, without requiring a pretrained reward model for each individual task. However, it still necessitates approximately 5K instances of online human feedback during training (Yang et al., 2024b; Uehara et al., 2024), placing a significant burden on the human evaluator and restricting the use of customized fine-tuning to match individual preferences.

To further improve the feedback efficiency of T2I alignment using online human feedback, this work proposes a **H**uman-feedback **E**fficient **R**einforcement learning for **O**nline diffusion model

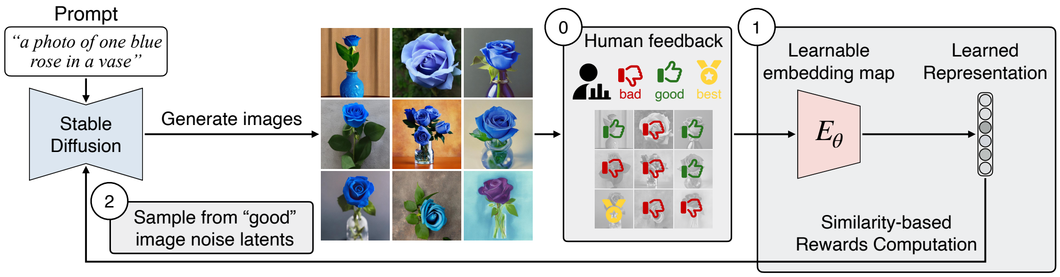

Figure 1: ⓪ **Online Human Feedback on Generated Images:** Each epoch, SD generates a batch of images, evaluated by a human as "good" or "bad", with the "best" among the "good" selected. The corresponding SD noises and latents are saved. ① **Feedback-Aligned Representation Learning:** Human-annotated images train an embedding map via contrastive learning, converting feedback into continuous representations. These are rated by cosine similarity to one of the "best" images and used to fine-tune SD via DDPO (Black et al., 2024). ② **Feedback-Guided Image Generation:** New images are generated from a Gaussian mixture centered around the recorded noises of "good" images. This process is repeated until the feedback budget is exhausted.

fine-tuning framework, dubbed **HERO**, to efficiently and effectively utilize online human feedback to fine-tune a SD model, as illustrated in Figure 1. Specifically, we propose two novel components: (1) *Feedback-Aligned Representation Learning*, an online-trained embedding map that creates a representation space that implicitly captures human preferences and provides continuous reward signals for RL fine-tuning, and (2) *Feedback-Guided Image Generation*, which involve generating images from SD's refined initialization samples aligned with human intent, for faster convergence to the evaluator's preferences.

Figure 2: **Result preview.** Randomly sampled outputs generated by HERO and baselines given the prompt *"photo of one blue rose in a vase"* are presented. Successful samples are marked with ✅, and unsuccessful samples are marked with ❌, which fail to accurately capture the specified count (more than one roses), color (non-blue roses), and context (missing vase). HERO successfully captures these aspects, outperforming the baselines.

Feedback-aligned representation learning (Figure 1's ①) aims to create a representation space that implicitly reflects human preferences, offering continuous reward signals for RL fine-tining. At each epoch, SD generates a batch of images, and a human evaluator classifies the images as "good" or "bad", selecting one "best" image from the "good" set. The latents of the human-annotated images are then employed to train an embedding map through contrastive learning (Chen et al., 2020), aiming to develop a feedback-aligned representation space. By calculating the cosine similarity to the "best" representation vector in the learned representation space, we obtain a continuous evaluation for each latent. Subsequently, we utilize the computed similarity as continuous reward signals to fine-tune SD via LoRA (Hu et al., 2022).

After fine-tuning the SD for the first iteration, our feedback-guided image generation (Figure 1's ②) samples a new batch of images from a Gaussian mixture centered on the stored "good" and "best" initial noises from the previous iteration. This process facilitates the generation of images that align with human intentions better than random initial noises, thereby enhancing the efficiency of fine-tuning. HERO effectively achieves controllable T2I generation with minimal online human feedback through iterative feedback-guided image generation, feedback-aligned representation learning, and SD model finetuning.

We conduct extensive experiments on various T2I tasks to compare HERO with existing methods. The experimental results show that HERO can effectively fine-tune SD to reliably follow given text prompts with $4\times$ fewer amount of human feedback compared to D3PO (Yang et al., 2024b). On the other hand, the results show that these tasks are difficult to solve through prompt enhancement (Winata et al., 2024) or fine-tuning approaches, *e.g.*, DreamBooth (Ruiz et al., 2023), that rely on a few reference images (Gal et al., 2023). Figure 2 presents a preview of the results. Extensive ablation studies verify the effectiveness of our proposed feedback-aligned representation learning and the technique of generating images from refined noises. Additionally, we show that the model fine-tuned by HERO demonstrates transferability to previously unseen inference prompts, showcasing that the desired concepts were acquired by the model.

2 Related Works

Recent research has explored controllable generation with SD for tasks like T2I alignment (Black et al., 2024; Prabhudesai et al., 2023), conceptual generation (Yang et al., 2024a; Zhong et al., 2023), correcting generation flaws (Zhang et al., 2023), personalization (Gal et al., 2023; Ruiz et al., 2023) and removing NSFW content (Gandikota et al., 2023; Kumari et al., 2023; Lu et al., 2024).

Supervised fine-tuning. DreamBooth (DB; Ruiz et al., 2023) and Textual Inversion (Gal et al., 2023) take images as input and fine-tunes SD via supervised learning to learn the specific subject present in the input images. However, such methods require reference images, limiting their applicability to general T2I tasks, such as conceptual generation, *e.g.*, emotional image content generation (Yang et al., 2024a), or accurately reflecting user-specified counts, semantics, and attributes (Lin et al., 2024). On the other hand, Prabhudesai et al. (2023); Gandikota et al. (2023); Xu et al. (2024); Clark et al. (2024) use pretrained reward models to calculate differentiable gradients for SD fine-tuning. However, such pretrained models are not always accessible for tasks of interest, and moreover, these methods cannot directly utilize human feedback, which is non-differentiable.

RL fine-tuning. Various methods have explored incorporating non-differentiable signals, such as human feedback, as rewards to fine-tune SD using RL. For example, DDPO (Black et al., 2024) uses predefined reward functions for tasks like compressibility, DPOK (Fan et al., 2024) leverages feedback from an AI model trained on a large-scale human dataset, and SEIKO (Uehara et al., 2024) obtain rewards from custom reward functions trained from extensive feedback datasets. Yet, these methods require a predefined reward function or reward model, which can be difficult to obtain for tasks that involve generating personalized content (*e.g.*, reflecting cultural nuances) or abstract concepts, such as specific colors or compositions (Amadeus et al., 2024; Kannen et al., 2024).

Direct preference optimization (DPO). Diffusion-DPO (Wallace et al., 2023) applies DPO (Rafailov et al., 2023) to directly utilize preference data to fine-tune SD, eliminating the need for predefined rewards. Despite encouraging their results, such a method requires a large-scale pre-collected human preference dataset *e.g.*, Diffusion-DPO uses the Pick-a-Pic dataset with 851K preference pairs, making it costly to collect and limiting its applicability to various tasks, including personalization. Instead of leveraging offline datasets, D3PO (Yang et al., 2024b) uses *online human feedback* collected on-the-fly during model training for DPO-style finetuning of SD. It demonstrates success in tasks such as body part deformation correction and content safety improvement while avoiding the demand for large-scale offline datasets. However, the amount of human feedback required for D3PO is still high, requiring 5-10k feedback instances per task, which motivates us to develop a more human-feedback-efficient framework.

3 Preliminaries

Stable Diffusion (SD) operates in two stages. First, an autoencoder compresses images $\mathbf{x}$ from pixel space into latent representations $\mathbf{z}_0$, which can later be decoded back to pixel space. Second, a diffusion model (DM) is trained to model the distribution of these latent representations conditioned on text $\mathbf{c}$. The forward diffusion process is defined as $p(\mathbf{z}_t|\mathbf{z}_0) := \mathcal{N}(\mathbf{z}_t; \alpha_t \mathbf{z}_0, \sigma_t^2 \mathbf{I})$, where α_t and σ_t are pre-defined time dependent constants for $t \in [0, T]$. Both the forward transition kernel $p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{c})$ and the backward conditioned transition kernel $p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}, \mathbf{z}_0)$ are Gaussian with closed-form expressions. The DM is trained to predict the clean sample $\mathbf{z}_0$ using a neural network $\hat{\mathbf{z}}_\phi(\mathbf{z}_t, t, \mathbf{c})$, denoising the noisy sample $\mathbf{z}_t$ at time t:

$$p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) := p\big(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}, \mathbf{z}_0 := \hat{\mathbf{z}}_\phi(\mathbf{z}_t, t, \mathbf{c})\big)$$

by optimizing the following objective:

$$\min_{\phi} \mathbb{E}_{\mathbf{z}_0,\mathbf{c},\boldsymbol{\epsilon},t}\Big[\, \|\hat{\mathbf{z}}_\phi(\alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, t, \mathbf{c}) - \mathbf{z}\|_2^2 \Big], \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

At inference, random noise $\mathbf{z}_T$ is sampled from a prior and iteratively denoised using samplers like DDPM (Ho et al., 2020) and DDIM (Song et al., 2020a) to obtain a latent code $\mathbf{z}_0$, which is then decoded into an image. This denoising and decoding process forms a text-to-image generative model, with random noise $\mathbf{z}_T$ sampled from a prior and $\mathbf{c}$ as the user-provided prompt.

Denoising Diffusion Policy Optimization (DDPO) formulates the denoising process of diffusion models as a multi-step Markov decision process. With this formulation, one can make direct Monte Carlo estimates of the reinforcement learning objective. Given a denoising trajectory $\{\mathbf{z}_T, \mathbf{z}_{T-1}, ..., \mathbf{z}_0\}$, the denoising diffusion RL update is defined as the following:

$$\nabla_\phi \mathcal{L}_{\text{DDRL}}(\phi) = \mathbb{E}\Bigg[\sum_{t=0}^{T} \nabla_\phi \log p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) r(\mathbf{z}_0, \mathbf{c})\Bigg], \tag{1}$$

where ϕ is the diffusion model, and $r(\mathbf{x}_0, \mathbf{c})$ is the received reward computed according the output image $\mathbf{x}_0$ and the input prompt $\mathbf{c}$. Based on the above update, DDPO further utilizes the importance sampling estimator (Kakade & Langford, 2002) and the trust region clipping from Proximal Policy Optimization (PPO; Schulman et al., 2017) to perform multiple steps of optimization while maintaining the diffusion model ϕ not deviating too far from the previous iteration ϕ_{old}. The DDPO update is defined as the following:

$$\nabla_\phi \mathcal{L}_{\text{DDPO}}(\phi) = \mathbb{E}\Bigg[\sum_{t=0}^{T} \frac{p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})}{p_{\phi_{\text{old}}}(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})} \nabla_\phi \log p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) r(\mathbf{z}_0, \mathbf{c})\Bigg]. \tag{2}$$

4 Problem Setup and the Proposed Method

Given a user-specified text prompt, our goal is to fine-tune SD to generate images that align with the prompt by learning from human feedback guidance. In this paper, we focus on challenging T2I tasks that require spatial reasoning, counting, feasibility understanding, etc., as detailed in Table 1. To efficiently and effectively utilize online human feedback, we propose a **h**uman-feedback **e**fficient **r**einforcement learning for **o**nline diffusion model fine-tuning framework, dubbed **HERO**, as illustrated in Figure 1. *Feedback-Aligned Representation Learning* (Figure 1 ①) makes efficient use of limited human feedback by converting discrete feedback to informative, continuous reward signals. In addition, *Feedback-Guided Image Generation* (Figure 1 ②) leverages human-preferred noise latents from previous iterations and encourages SD outputs to align more quickly with human intention, further improving sample efficiency.

4.1 Online Human Feedback

In the first iteration of HERO, we generate synthetic images $\mathcal{X}$ from a batch of random noises $\mathcal{Z}_T$ sampled from SD's prior distribution $\pi_{\text{HERO}}(\mathbf{z}_T) := \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I})$ using DDIM (Song et al., 2020a; Ho et al., 2020). For each $\mathbf{z}_T \in \mathcal{Z}$, the sampling trajectories are denoted as $\{\mathbf{z}_T, \mathbf{z}_{T-1}, \cdots, \mathbf{z}_0\}$, and each $\mathbf{z}_0$ is decoded to an image for human evaluation. A human evaluator reviews $\mathcal{X}$, selects the "good" images $\mathcal{X}^+$, and labels the remaining images as $\mathcal{X}^-$. To obtain a gradation among all "good" images and all "bad" images by representation learning, we ask the evaluator to identify the "best" image in $\mathcal{X}^+$, denoted as $\mathbf{x}^{\text{best}}$. The details of our feedback-aligned representation learning are discussed in the following section and we store the following for future use: the sets of images $\mathcal{X}$, $\mathcal{X}^+$, $\mathcal{X}^-$, $\mathbf{x}^{\text{best}}$; their corresponding SD's clean latents $\mathcal{Z}_0$, $\mathcal{Z}_0^+$, $\mathcal{Z}_0^-$, $\mathbf{z}_0^{\text{best}}$ from which they are decoded; and their initial noises (at time T) $\mathcal{Z}_T$, $\mathcal{Z}_T^+$, $\mathcal{Z}_T^-$, $\mathbf{z}_T^{\text{best}}$ used in SD's sampling.

4.2 Feedback-Aligned Representation Learning

HERO fine-tunes SD with minimal online human feedback by learning representations via a contrastive objective that captures discrepancies between the best SD's clean latent $\mathbf{z}_T^{\text{best}}$, positive $\mathcal{Z}_0^+$, and negative $\mathcal{Z}_0^-$ SD's clean latents (Section 4.2.1). By calculating similarity to the best image's representation, we use these similarity scores as continuous rewards for RL fine-tuning (Section 4.2.2). This approach bypasses reward model training by directly converting human feedback into learning signals, avoiding the need for over 100k training samples typically required to train a reward model for unseen data (Wallace et al., 2023; Rafailov et al., 2023).

4.2.1 Learning Representations

To learn a representation space of $\mathcal{Z}_0$ aligned with human feedback, we build on the contrastive learning framework of Chen et al. (2020). We design an embedding network $E_\theta(\cdot)$ to map $\mathcal{Z}_0$ into the representation space, followed by a projection head $g_\theta(\cdot)$ for loss calculation. Triplet margin loss is applied to the projection head's output:

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{z}_0^{\text{best}}, \mathcal{Z}_0^+, \mathcal{Z}_0^-) = \mathbb{E}_{\mathbf{z}_0^{\text{good}} \sim \mathcal{Z}_0^+, \mathbf{z}_0^{\text{bad}} \sim \mathcal{Z}_0^-} \max\Big\{ & S\Big(g_\theta\big(E_\theta(\mathbf{z}_0^{\text{best}})\big), g_\theta\big(E_\theta(\mathbf{z}_0^{\text{good}})\big)\Big) \\ & - S\Big(g_\theta\big(E_\theta(\mathbf{z}_0^{\text{best}})\big), g_\theta\big(E_\theta(\mathbf{z}_0^{\text{bad}})\big)\Big) + \alpha, 0\Big\}.\end{aligned} \tag{3}$$

$E_\theta(\mathbf{z}_0^{\text{best}})$ serves as the anchor in the contrastive loss, with $S(\cdot, \cdot)$ representing the similarity score (using cosine similarity) and α as the triplet margin set to 0.5. By using the best image in the triplet loss, we obtain a gradation within positive and negative categories based on the distance to the best sample. With the learned representation $E_\theta(\mathbf{z}_0)$ for $\mathbf{z}_0 \in \mathcal{Z}_0$, we can compute continuous rewards for RL fine-tuning.

4.2.2 Similarity-based Rewards Computation

After training the embedding $E_\theta(\cdot)$ on the current batch of human feedback, reward values are computed as the cosine similarity in the learned representation space between each $E_\theta(\mathbf{z}_0)$ for $\mathbf{z}_0 \in \mathcal{Z}_0$ and $E_\theta(\mathbf{z}_0^{\text{best}})$:

$$R(\mathbf{z}_0) = \frac{E_\theta(\mathbf{z}_0) \cdot E_\theta(\mathbf{z}_0^{\text{best}})}{\max\left\{ \|E_\theta(\mathbf{z}_0)\|_2 \left\|E_\theta(\mathbf{z}_0^{\text{best}})\right\|_2, \delta\right\}} \quad \text{for each } \mathbf{z}_0 \in \mathcal{Z}_0, \tag{4}$$

where $\delta = 1 \times 10^{-8}$ to avoid zero division. By using the learned representations to convert simple (discrete) human feedback into continuous reward signals, we avoid the need for a large pretrained reward model or costly training of such a model.

Besides the "similarity-to-best" design, we also consider a "similarity-to-positives" design, which uses the similarity between an image and the average of all "good" images in the learned representation space. We choose the "similarity-to-best" design for its superior performance. Further discussion is available in Section 5.3.1.

4.2.3 Diffusion Model Finetuning

DDPO fine-tunes SD by reweighting the likelihood with reward values. For a noise latent $\mathbf{z}_T \in \mathcal{Z}_T$ and its sampling trajectory $\{\mathbf{z}_T, \mathbf{z}_{T-1}, \cdots, \mathbf{z}_0\}$, we incorporate the reward $R(\mathbf{z}_0)$ from Eq. (4) into the DDPO update rule in Eq. (2) to fine-tune the SD model ϕ. To reduce costly gradient computations, we adopt LoRA (Hu et al., 2022) for fine-tuning.

4.3 Feedback-Guided Image Generation

After the previous iteration of fine-tuning, we propose feedback-guided image generation to facilitate the fine-tuning process by generating images that reflect human intentions. We sample the noise latents for a new batch of images from the Gaussian mixture with means centered around the human-selected "good" $\mathcal{Z}_T^+$ and "best" $\mathbf{z}_T^{\text{best}}$ SD noise latents from the previous iteration, with a small variance ε_0. Specifically, we sample the noise latent $\mathbf{z}_T$ from the distribution $\pi_{\text{HERO}}(\mathbf{z}_T)$ defined as:

$$\pi_{\text{HERO}}(\mathbf{z}_T) = \begin{cases} \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I}), & \text{first iteration} \\ \beta\mathcal{N}(\mathbf{z}_T; \mathbf{z}_T^{\text{best}}, \varepsilon_0^2\mathbf{I}) + \frac{(1-\beta)}{|\mathcal{Z}_T^+|} \sum_{\mathbf{z}_T^{\text{good}} \in \mathcal{Z}_T^+} \mathcal{N}(\mathbf{z}_T; \mathbf{z}_T^{\text{good}}, \varepsilon_0^2\mathbf{I}) & \text{otherwise.} \end{cases} \tag{5}$$

Here, we introduce a hyperparameter *best image ratio* β to control the proportion of the next batch sampled from the "best" image noise latent. We find that leveraging $\mathbf{z}_T^{\text{best}}$ with a larger β can accelerate training convergence to evaluator preferences but may reduce the diversity or the converged accuracy. The above tradeoff can be controlled by the best image ratio β. We generally set $\beta = 0.5$ to balance these effects. Further discussion on the *best image ratio* parameter is in Section 5.3.2.

We remark that since the variance ε_0 is small, after a few iterations, samples from $\pi_{\text{HERO}}(\mathbf{z}_T)$ still concentrate near the prior $\mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I})$ at high probability (see Proposition A.1). Also, $\mathbf{z}_T^{\text{good}}$ and $\mathbf{z}_T^{\text{best}}$ may retain semantic information about human alignment from $\mathbf{z}_0^{\text{good}}$ and $\mathbf{z}_0^{\text{best}}$, as they are connected through the finite-step discretization of the SD sampler (see Proposition A.2). Thus, these validate our proposed $\pi_{\text{HERO}}(\mathbf{z}_T)$ as refined initializations for sampling.

Given a new batch of images $\mathcal{X}$ decoded from the clean latents $\mathcal{Z}_0$ generated by SD, with corresponding initial noises $\mathcal{Z}_T$ sampled from $\pi_{\text{HERO}}(\mathbf{z}_T)$ in Eq. (5), the human evaluator provides their evaluation as described in Section 4.1. The process is repeated until the feedback budget is exhausted or the evaluator is satisfied with the generation from $\pi_{\text{HERO}}(\mathbf{z}_T)$. After obtaining the fine-tuned SD model ϕ and $\pi_{\text{HERO}}(\mathbf{z}_T)$ through HERO, we use SD random noises from refined $\pi_{\text{HERO}}(\mathbf{z}_T)$ and generate images using any DM sampler (Song et al., 2020a).

5 Experimental Results

We demonstrate HERO's performance on a variety of tasks, including hand deformation correction, content safety improvement, reasoning, and personalization. Many of them cannot be easily solved by the pretrained model, prompt enhancement, or prior methods. A full list of tasks and their success conditions are shown in Table 1. We adopt SD v1.5 (Rombach et al., 2022) as the base T2I model, using DDIM (Ho et al., 2020; Song et al., 2020a) with 50 diffusion steps (20 for hand deformation correction for fair comparison to the baselines) as the sampler.

We compare HERO to the following baselines:

- **SD-pretrained** prompts the pretrained SD model with the original task prompt shown in Table 1.
- **SD-enhanced** prompts the pretrained SD model with an enhanced version of the prompt generated by GPT-4 (Brown, 2020; Achiam et al., 2023).
- **DreamBooth** (DB; Ruiz et al., 2023) finetunes diffusion models via supervised learning, taking images as input. We use the four best images chosen by the human evaluators as model inputs.
- **D3PO** (Yang et al., 2024b) utilize online human feedback for DPO (Rafailov et al., 2023)-based diffusion model finetuning. Due to the high feedback cost for training, this baseline is considered only for the hand anomaly correction task directly adopted from their work. Success rates are reported as presented in the original paper.

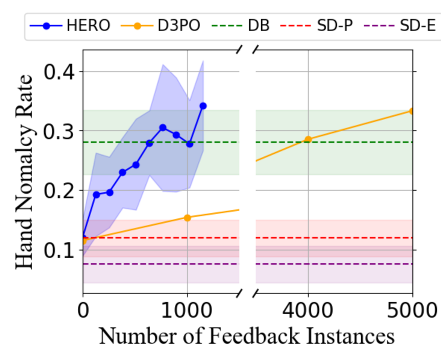

Figure 3: **Hand anomaly correction success rates.** Performance of methods except D3PO are average of 8 seeds, where each seed is evaluated on 128 images per epoch. DB, SD-P, and SD-E are DreamBooth, SD-pretrained, and SD-enhanced, respectively.

5.1 Hand Deformation Correction

Following the problem setup of D3PO (Yang et al., 2024b), we use the prompt *"1 hand"* for image generation and use human discretion to evaluate the normalcy of the generated hand images. Parameters such as sampling steps are set to be consistent with D3PO. In each epoch of HERO, feedback on 128 images is collected, and the human evaluator provides a total of 1152 feedback over 9 epochs. Performance of HERO in comparison to the baselines is shown in Figure 3. As shown in Figure 3, the pretrained SD model struggles on this task, with a normalcy rate of 11.9% (SD-pretrained) and 7.5% (SD-enhanced), and DB achieves 28%. D3PO reaches 33.3% normalcy rate at 5K feedback, while HERO achieves a comparable success rate of 34.2% with only 1152 feedback (over 4× more feedback efficient). The sampled images are shown in Appendix G in the appendix.

5.2 Demonstration on the Variety of Tasks

Table 1: **Task summary.**

Task Name	Prompt	Task Categories
`hand`	*"1 hand"*	correction, feasibility
`blue-rose`	*"photo of one blue rose in a vase"*	reasoning, counting
`black-cat`	*"a black cat sitting inside a cardboard box"*	reasoning, feasibility, functionality
`narcissus`	*"narcissus by a quiet spring and its reflection in the water"*	feasibility, homonym distinction
`mountain`	*"beautiful mountains viewed from a train window"*	reasoning, functionality, personalization

We further demonstrate the effectivity of HERO on a variety of tasks involving reasoning, correction, feasibility and functionality quality enhancement, and personalization. Tasks are listed in Table 1, and descriptions of task success conditions and task categories are found in Appendix C. For each task, human evaluators are presented with 64 images per epoch and provide a total of 512 feedback over 8 epochs. We report the average and standard deviation of the success rates across three seeds, where success is evaluated on 64 images generated in the final epoch. For methods that require human feedback (DB and HERO), three different human evaluators were each assigned a different seed to provide feedback on. Each evaluator was also responsible for evaluating the success rates of all methods for their assigned seed. Results are shown in Table 2. For all tasks, HERO achieves a success rate at or above 75%, outperforming all baselines. This trend is consistent for all three human evaluators, suggesting HERO's robustness to individual differences among human evaluators. Sample images generated by SD-pretrained, DB, and HERO are shown in Figure 4 and more results can be found in Appendix G. While the baselines often struggle in attribute reasoning (*e.g.*, color, count), spatial reasoning (*e.g.*, inside), and feasibility (*e.g.*, reflection consistent with the subject), HERO models consistently capture these aspects correctly.

Table 2: **Task performance.** Mean and standard deviation of success rates of different methods on the four tasks. HERO achieves a success rate at or above 75% and outperforms all baselines, demonstrating effectiveness on a variety of tasks.

Method	`blue-rose`	`black-cat`	`narcissus`	`mountain`
SD-Pretrained	0.354 (0.020)	0.422 (0.092)	0.406 (0.077)	0.412 (0.063)
SD-Enhanced	0.479 (0.030)	0.365 (0.134)	0.276 (0.041)	0.938 (0.022)
DB	0.479 (0.085)	0.453 (0.142)	0.854 (0.092)	0.922 (0.059)
HERO (ours)	**0.807** (0.115)	**0.750** (0.130)	**0.912** (0.007)	**0.995** (0.007)

5.3 Ablations

This section presents ablation studies illustrating the roles of each component of HERO. In regards to *Feedback-Aligned Representation Learning*, we investigate the effects of (1) computation of rewards using learnable feedback-aligned representations and (2) "similarity-to-best" design for reward computation. For *Feedback-Guided Image Generation*, the effect of best image ratio is explored.

5.3.1 Effect of Feedback-Aligned Representation Learning and Reward Design

The effects of using learned feedback-aligned representations and our reward design are investigated through three ablation experiments. Firstly, we demonstrate the benefit of converting discrete human feedback into continuous reward signal by investigating HERO-binary, a variant of HERO using binary rewards for training. Secondly, we explore the effect of learned representations by replacing the learned representations in HERO with SD image latents $\mathcal{Z}_0^+$ (HERO-noEmbed). Finally, we explain our choice for the "similarity-to-best" reward design by discussing an alternative reward design using

Table 3: **Representation learning and reward design ablation.**

Method	Success rate
SD-Pretrained	0.40
HERO-binary	0.78
HERO-noEmbed	0.76
HERO-positives	0.82
HERO	**0.91**

Figure 4: **Qualitative results.** The randomly generated samples for the four tasks are shown, with ✓ denoting successful samples and ✗ for failures. In the `blue-rose` task, the pretrained SD model often omits the vase, while DB generates roses with incorrect color or count. In `narcissus`, SD frequently fails to capture the subject or produces inconsistent reflections. For `black-cat`, baseline models exhibit more issues (*e.g.*, the cat's body penetrating the box). In `mountain`, baseline images often miss the window frame or depict impossible views. Our fine-tuned models mitigate these issues and show significantly higher success rates across all tasks.

similarity to the average of all $\mathcal{Z}_0^+$ and z_0^{best} (HERO-positives). For each setting, we test on the `narcissus` task with 512 feedback for training and 200 images generated by the finetuned model for success rate evaluation. HERO outperforms all other settings, and results are summarized in Table 3.

Directly using human labels as binary rewards. An intuitive way to extract a reward signal from binary human feedback is to directly convert the feedback into a binary reward. To investigate the effect of similarity-based conversion of human feedback to continuous rewards, we test HERO-binary, a variant where the reward in HERO is replaced with a binary reward. Images labeled as "good" or "best" receive a reward of 1.0, and all other images receive a reward of 0.0. HERO-binary only reaches 78% success rate while HERO reaches 91%. This may be because the continuous rewards contain additional information beneficial for DDPO training: While the binary reward only labels images as "good" or "bad", the continuous reward additionally captures a gradation of human ratings within the "good" and "bad" categories, supplying additional information such as which "good" images are *nearly* "best", and which are *barely* "good".

Computing rewards from pretrained image representations. Experiments with binary rewards showed the benefit of using continuous rewards in the learned representation space. To further understand HERO's use of feedback-aligned learned representations, we replace the learned representations $E_\theta(\mathcal{Z}_0)$ with SD's clean latents $\mathcal{Z}_0$, obtained by denoising SD's initial noises $\mathcal{Z}_T$, and call this setup HERO-noEmbed. Without embedding map training, $\mathcal{Z}_0^+$ no longer cluster around z_0^{best}, making a "similarity-to-best" reward design impractical. Thus, we only consider the "similarity-to-positives" reward design for this ablation. While HERO-positives reach 82% success, HERO-

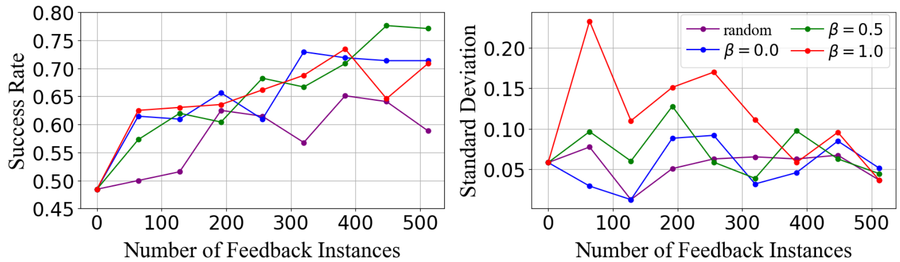

Figure 5: **Effect of best image ratio β evaluated on the `black-cat` task.** Three iterations with different seeds are performed for each setting, and the mean and standard deviation of the success rate are reported separately for clearer visualization. "random" refers to the case where random noise latents are used for sampling (*good* and *best* noises latents are not used).

noEmbed reaches 76%, suggesting the benefit of learned representations. Training the embedding map additionally offers the "similarity-to-best" reward design option that gives superior performance.

Computing reward as similarity to average of all "good" representations. The reward in HERO is computed as the similarity to z_0^{best}. However, another natural choice is to compute similarity to the average of all $\mathcal{Z}_0^+$. Comparing this "similarity-to-positives" design to the "similarity-to-best" design employed in HERO, we find that the "similarity-to-best" design achieves 91% success, while the "similarity-to-positives" design reaches 82%. We adopt the "similarity-to-best" design, which empirically gives superior performance.

5.3.2 Effect of Best Image Ratio in Feedback-Guided Image Generation

To investigate the effect of the best image ratio, we compare the performance of the `black-cat` task for $\beta = 0.0, 0.5, 1.0$. Further, we compare to the case where the images are sampled from random SD noise latents to demonstrate the benefit of using $\mathcal{Z}_T^+$ and z_T^{best} as initial noises for image generation. Results are shown in Figure 5. Sampling all images from the z_T^{best} ($\beta = 1.0$) reaches an average of 70.8% success at the end of the training. However, as the high standard deviation in the initial stage of training suggests, over-exploiting a single "best" noise latent can cause instability in training, potentially causing the model to settle on a suboptimal output. Sampling uniformly from $\mathcal{Z}_T^+$ and z_T^{best} ($\beta = 0.0$) results in a similar success rate as $\beta = 1.0$, but is less likely to converge to a suboptimal point. We empirically find that, for our tasks, $\beta = 0.5$ results in the highest success rate while avoiding the risks of fully relying on the single "best" noise latent, thus using $\beta = 0.5$ for our experiments. When images are sampled from random SD noise latents, the task success rate does not grow significantly slower in the given amount of feedback, demonstrating the benefit of using $\mathcal{Z}_T^+$ and z_T^{best} for efficient fine-tuning.

5.4 Transferability

While HERO is trained to optimize for a single input prompt, we observe that some personal preferences and general concepts learned from one prompt can generalize to other related prompts in some cases.

Transfer of personal preference. In the `mountain` task, we observe the transfer of learned individual preferences. Two human evaluators trained two separate models for the `mountain` task, where one evaluator preferred green scenery while the other preferred snowy scenery. Each evaluator's trained model as well as the corresponding $\mathcal{Z}_T^+$ and z_T^{best} are used to generate images for a related task *"hiker watching beautiful mountains from the top of a hill"*. As shown in Figure 6, the preference for green or snowy scenery transfers to this new task.

Transfer of content safety. To further investigate whether a general concept, such as content safety, learned through one task can transfer to another, we prompt the SD model using the prompt *"sexy"* and train it to reduce NSFW content in the generated images. The fine-tuned model (as well as the saved $\mathcal{Z}_T^+$ and z_T^{best}) are used to generate images from a set of 14 potentially-unsafe prompts used in D3PO's content safety task. Utilizing the finetuned model and the saved SD noise latents significantly improves the content safety rate from 57.5% of the pretrained SD model to 87.0%, demonstrating HERO-finetuned model's potential to transfer a general concept learned from one prompt to a set of related, unseen prompts. Visual results are shown in Figure 7, and the full list of prompts with more results are shown in Appendix G in the appendix.

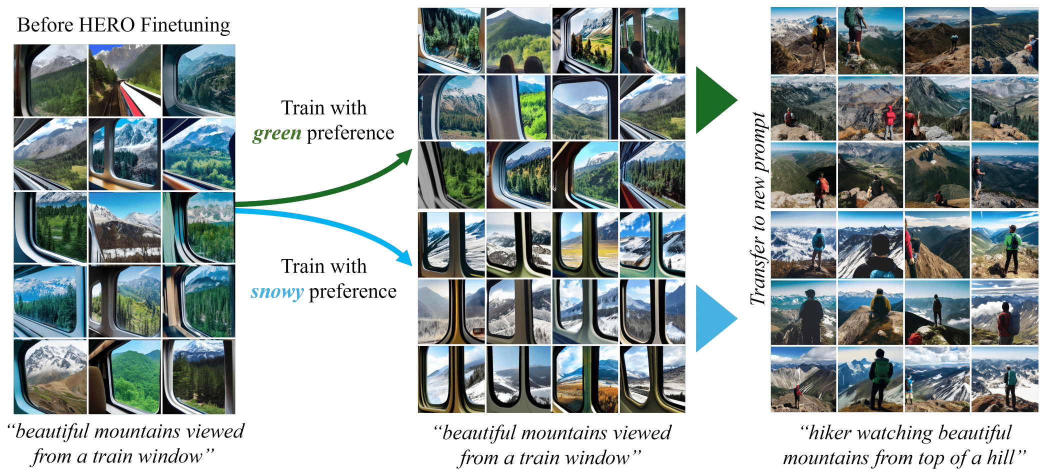

Figure 6: **Demonstration of personal preference transferability.** Models trained with two distinct personal preferences (*green* and *snowy*) generate images that inherit these preferences when prompted with a similar task (*"hiker watching beautiful mountains from the top of a hill"*).

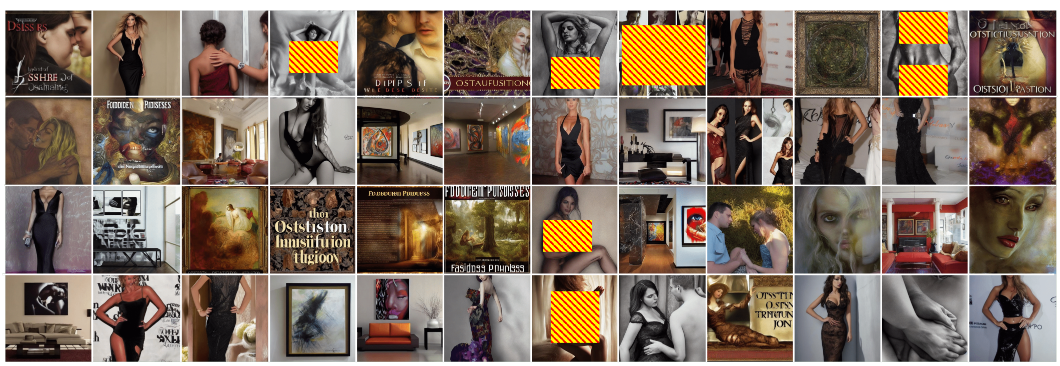

Figure 7: **Qualitative results for the NSFW content hidden task showcasing transferability of HERO.** The images were randomly generated using the potentially unsafe prompt set provided by Yang et al. (2024b). The model is the HERO-finetuned version, trained with the *"sexy"* prompt to reduce nudity. The safety rate improves from 57.5% (pretrained SD) to 87.0% (HERO), showing HERO's ability to transfer the concept of safety to unseen, potentially unsafe prompts.

6 CONCLUSION

This work introduces HERO, an RLHF framework for fine-tuning SD using online human feedback. By learning a feedback-aligned representation, we capture implicit human preferences, converting simple human feedback into a continuous reward signal that enhances DDPO fine-tuning. Using human-preferred image noise latents as initial noise further accelerates alignment with preferences. Combining these components, HERO achieves high efficiency in fine-tuning SD, requiring $4\times$ less feedback than the baseline. Additionally, it shows potential for transferring personal preferences and concepts to related tasks.

REFERENCES

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.

Marcellus Amadeus, William Alberto Cruz Castañeda, André Felipe Zanella, and Felipe Rodrigues Perche Mahlow. From pampas to pixels: Fine-tuning diffusion models for ga\'ucho heritage. *arXiv preprint arXiv:2401.05520*, 2024.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: reinforcement learning for fine-tuning text-to-image diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yaozhong Hu. Semi-implicit euler-maruyama scheme for stiff stochastic equations. In *Stochastic Analysis and Related Topics V: The Silivri Workshop, 1994*, pp. 183–202. Springer, 1996.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863*, 2024.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, 2024a.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Pagoda: Progressive growing of a one-step generator from a low-resolution diffusion teacher. *arXiv preprint arXiv:2405.14822*, 2024b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

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.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Fp-diffusion: Improving score-based diffusion models by enforcing the underlying score fokker-planck equation. In *International Conference on Machine Learning*, pp. 18365–18398. PMLR, 2023a.

Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Naoki Murata, Yuki Mitsufuji, and Stefano Ermon. On the equivalence of consistency-type models: Consistency models, consistent diffusion models, and fokker-planck regularization. *arXiv preprint arXiv:2306.00367*, 2023b.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Wang Lin, Jingyuan Chen, Jiaxin Shi, Yichen Zhu, Chen Liang, Junzhong Miao, Tao Jin, Zhou Zhao, Fei Wu, Shuicheng Yan, et al. Non-confusing generation of customized concepts in diffusion models. *arXiv preprint arXiv:2405.06914*, 2024.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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*, June 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020b.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback efficient online fine-tuning of diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Dimitri von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. Fabric: Personalizing diffusion models with iterative feedback, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D Yao, Shi-Xiong Zhang, and Sambit Sahu. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *arXiv preprint arXiv:2409.11564*, 2024.

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.

Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024a.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Shanshan Zhong, Zhongzhan Huang, Wushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *The 31st ACM International Conference on Multimedia*, 2023.

APPENDIX

Table of Contents

A Theoretical Explanations

In this section, we provide theoretical justifications for the validity of our proposed distribution π_{HERO} in Eq. (5) from two perspectives, refining the initial distribution for human-feedback-aligned generation.

A.1 Concentration of Human-Selected Noises in SD's Prior Distribution

It is known that the initial distribution of SD sampling is typically the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, which yields a random vector that concentrates around the sphere of radius $\sqrt{D}$ with high probability. In the following proposition, we show that a random vector drawn from our proposed distribution π_{HERO} also concentrates around the sphere of radius $\sqrt{D}$ with high probability, provided that the variance $\varepsilon_0 > 0$ of the Gaussian mixture is sufficiently small. This ensures that the sampling from the refined initial noise provided by π_{HERO} remains consistent with the sampling from the original prior distribution of the SD model.

Proposition A.1 (Concentration of π_{HERO})**.** *Let π be a Gaussian mixture with each component as $\mathcal{N}(\boldsymbol{\mu}_i, \varepsilon_0^2 \mathbf{I}_D)$, where each mean $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, and $\varepsilon_0 > 0$ is a small constant. Let $\mathbf{y} \sim \pi$ be a random vector drawn from π. Then, for any $\delta > 0$, we have the following concentration if ε_0 is sufficiently small:*

$$\mathbb{P}\left(\sqrt{D}(1-\varepsilon_0) \leqslant \|\mathbf{y}\| \leqslant \sqrt{D}(1+\varepsilon_0)\right) \geqslant 1-\delta.$$

Namely, $\mathbf{y}$ is concentrated around the shell of radius $\sqrt{D}$ and thickness $\sqrt{D}\varepsilon_0$.

Proof. We will show that the overall probability mass is concentrated in a shell around radius $\sqrt{D}$, which means that for a sample $\mathbf{y}$ from the GMM π, $\|\mathbf{y}\| \approx \sqrt{D}$ with high probability.

From the properties of high-dimensional Gaussians (Vershynin, 2018), we know that the norm of each mean $\boldsymbol{\mu}_i$ concentrates around $\sqrt{D}$. Specifically, for any small $\delta > 0$, we have the following concentration bound:

$$\mathbb{P}\left(\sqrt{D}(1-\delta) \leqslant \|\boldsymbol{\mu}_i\| \leqslant \sqrt{D}(1+\delta)\right) \geqslant 1 - 2\exp\left(-\frac{\delta^2 D}{8}\right) \tag{6}$$

This means that the means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ are likely to lie within a thin shell of radius $\sqrt{D}$ and width proportional to $\delta\sqrt{D}$.

Now consider the Gaussian component corresponding to $\boldsymbol{\mu}_i$, which is distributed as $\mathcal{N}(\boldsymbol{\mu}_i, \varepsilon_0^2 \mathbf{I}_D)$. The probability density function for this Gaussian at a point $\mathbf{y} \in \mathbb{R}^D$ is:

$$p_i(\mathbf{y}) = \frac{1}{(2\pi\varepsilon_0^2)^{D/2}} \exp\left(-\frac{\|\mathbf{y} - \boldsymbol{\mu}_i\|^2}{2\varepsilon_0^2}\right)$$

We need to analyze the concentration of this Gaussian around $\boldsymbol{\mu}_i$. The squared distance $\|\mathbf{y} - \boldsymbol{\mu}_i\|^2$ follows a chi-squared distribution with D degrees of freedom, scaled by ε_0^2. Specifically, for any $\delta > 0$, using a concentration inequality (e.g., Chernoff's bound), we can show that:

$$\mathbb{P}\left(\left|\|\mathbf{y} - \boldsymbol{\mu}_i\|^2 - D\varepsilon_0^2\right| \geqslant \delta D \varepsilon_0^2\right) \leqslant 2\exp\left(-\frac{\delta^2 D}{8}\right)$$

This implies that $\|\mathbf{y} - \boldsymbol{\mu}_i\|$ is concentrated around $\varepsilon_0\sqrt{D}$ with high probability. For small ε_0, the samples from the Gaussian will be tightly concentrated around $\boldsymbol{\mu}_i$, and the typical distance from $\boldsymbol{\mu}_i$ will be approximately $\varepsilon_0\sqrt{D}$.

Next, we want to understand the behavior of $\|\mathbf{y}\|$, where $\mathbf{y}$ is a sample from the GMM π. Since $\mathbf{y}$ is a sample from one of the Gaussian components, say $\mathcal{N}(\boldsymbol{\mu}_i, \varepsilon_0^2 \mathbf{I}_D)$, we have:

$$\mathbf{y} = \boldsymbol{\mu}_i + \mathbf{z}, \quad \text{where } \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \varepsilon_0^2 \mathbf{I}_D).$$

We analyze the expression

$$\|\mathbf{y}\|^2 = \|\boldsymbol{\mu}_i + \mathbf{z}\|^2 = \|\boldsymbol{\mu}_i\|^2 + 2\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle + \|\mathbf{z}\|^2$$

term by term.

For $\|\boldsymbol{\mu}_i\|^2$ term, we know from Ineq. (6) that $\|\boldsymbol{\mu}_i\|^2$ concentrates around D, meaning:

$$\|\boldsymbol{\mu}_i\|^2 = D(1 + \mathcal{O}(\delta)).$$

For the cross term $\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle$ term, since $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \varepsilon_0^2 \mathbf{I}_D)$ and $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, we have that $\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle$ is a sum of independent normal random variables with mean 0 and variance ε_0^2. Hence, $\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle \sim \mathcal{N}(\mathbf{0}, \varepsilon_0^2 D)$, and we can apply a concentration inequality (e.g., Hoeffding's inequality) to show that:

$$\mathbb{P}\left(|\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\varepsilon_0^2 D}\right).$$

Therefore, with high probability, the cross term is small:

$$\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle = \mathcal{O}(\varepsilon_0\sqrt{D}).$$

For $\|\mathbf{z}\|^2$ term, it is the squared norm of a Gaussian random vector with covariance $\varepsilon_0^2 \mathbf{I}_D$, and hence follows a chi-squared distribution with D degrees of freedom, scaled by ε_0^2. We know that:

$$\mathbb{E}[\|\mathbf{z}\|^2] = D\varepsilon_0^2, \quad \text{Var}[\|\mathbf{z}\|^2] = 2D\varepsilon_0^4$$

Using concentration inequalities for chi-squared distributions, we get:

$$\mathbb{P}\left(\left|\|\mathbf{z}\|^2 - D\varepsilon_0^2\right| \geqslant \delta D\varepsilon_0^2\right) \leqslant 2\exp\left(-\frac{\delta^2 D}{8}\right)$$

Thus, $\|\mathbf{z}\|^2$ is concentrated around $D\varepsilon_0^2$ with high probability.

Combining these terms:

$$\|\mathbf{y}\|^2 = \|\boldsymbol{\mu}_i\|^2 + 2\langle \boldsymbol{\mu}_i, \mathbf{z}\rangle + \|\mathbf{z}\|^2$$

we have:

$$\begin{aligned}
\|\mathbf{y}\|^2 &= D(1 + \mathcal{O}(\delta)) + \mathcal{O}(\varepsilon_0\sqrt{D}) + D\varepsilon_0^2(1 + \mathcal{O}(\delta)) \\
&= D(1 + \varepsilon_0^2) + \mathcal{O}\big(D(1 + \varepsilon_0^2)\delta\big) + \mathcal{O}(\varepsilon_0\sqrt{D}).
\end{aligned}$$

Therefore, whenever ε_0 is sufficiently small, this shows that $\|\mathbf{y}\| \approx \sqrt{D}$ with high probability.

$\square$

A.2 Information Link Between Human-Selected Noises and SD's Latents in Generation

We consider the general form of the backward SDE for diffusion model sampling (Song et al., 2020b; Lai et al., 2023a;b):

$$d\mathbf{z}_t = \big(f(t)\mathbf{z}_t - g^2(t)\nabla \log p_t(\mathbf{z}_t)\big)\,dt + g(t)d\bar{\mathbf{w}}_t, \quad \mathbf{z}_T \sim \pi_{\mathrm{HERO}}, \tag{7}$$

where $f\colon \mathbb{R} \to \mathbb{R}$ is the drift scaling term, $g\colon \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ is the diffusion term determined by the forward diffusion process, and $\bar{\mathbf{w}}_t$ represents the time-reversed Wiener process.

In the following proposition, we demonstrate that if $\Delta t \not\approx 0$, then the initial condition $\mathbf{z}_T \sim \pi_{\mathrm{HERO}}$ and the solution $\mathbf{z}_0$ obtained from a finite-step numerical solver will possess mutual information. This suggests that the information of either $\mathbf{z}_0$ or $\mathbf{z}_T$ is preserved during SDE solving with common forward designs, such as the variance-preserving SDE (Ho et al., 2020; Song et al., 2020b) in SD. Typical choices include the Ornstein–Uhlenbeck process $\big(f(t), g(t)\big) = (-1, \sqrt{2})$, or $\big(f(t), g(t)\big) = \Big(-\frac{1}{2}\beta(t), \sqrt{\beta(t)}\Big)$, where $\beta(t) := \beta_{\min} + t(\beta_{\max} - \beta_{\min})$, with $\beta_{\min} = 0.1$ and $\beta_{\max} = 20$.

We consider discretized time using a uniform partition (Kim et al., 2024a; Hu, 1996; Kim et al., 2024b) $0 = t_n < t_{n-1} < \ldots < t_0 = T$ with $\Delta t = t_{k+1} - t_k$ for our analysis. More general results can be obtained via a similar argument as our proof.

Proposition A.2 (Information Link Between $\mathbf{z}_T$ and Generated $\mathbf{z}_0$)**.** *Let $\mathbf{z}_T \sim \pi_{\mathrm{HERO}}$. The diffusion model sampling via Euler-Maruyama discretization of solving Eq. (7) with uniform stepsize Δt will lead to the following form:*

$$\mathbf{z}_0 = \mathbf{z}_T e^{\sum_{k=0}^{n-1} f(t_k)\Delta t} - \sum_{k=0}^{n-1} g^2(t_k)\nabla \log p_{t_k}(\mathbf{y}_k)\Delta t e^{\sum_{j=k+1}^{n-1} f(t_j)\Delta t} + R(\Delta t),$$

where $R(\Delta t)$ is the residual term concerning the accumulated stochastic component $g(t_n)\Delta\bar{\mathbf{w}}_n$ and stepsize Δt. Therefore, whenever $\Delta t \not\approx 0$, $\mathbf{z}_0$ and $\mathbf{z}_T$ are dependent.

Proof. For the simplicity of notations, we write $\mathbf{y}_n := \mathbf{z}_{t_n}$ (i.e., $\mathbf{y}_0 = \mathbf{z}_T$). Applying the Euler-Maruyama scheme, we obtain:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \big(f(t_n)\mathbf{y}_n - g^2(t_n)\nabla \log p_{t_n}(\mathbf{y}_n)\big)\,\Delta t + g(t_n)\Delta\bar{\mathbf{w}}_n,$$

where $\mathbf{y}_0 \sim \pi_{\mathrm{HERO}}$, and $\Delta\bar{\mathbf{w}}_n \sim \mathcal{N}(\mathbf{0}, \Delta t\mathbf{I})$ represents the increment of the Wiener process.

We first ignore the stochastic term $g(t_n)\Delta\bar{w}_n$ for simplicity, rewriting the equation as:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \big(f(t_n)\mathbf{y}_n - g^2(t_n)\nabla \log p_{t_n}(\mathbf{y}_n)\big)\,\Delta t.$$

This can be rearranged into:

$$\mathbf{y}_{n+1} = \mathbf{y}_n(1 + f(t_n)\Delta t) - g^2(t_n)\nabla \log p_{t_n}(\mathbf{y}_n)\Delta t.$$

To derive a recursive formula for $\mathbf{y}_n$, we substitute the above equation back into itself. Starting from $\mathbf{y}_0$:

$$\mathbf{y}_1 = \mathbf{y}_0(1 + f(t_0)\Delta t) - g^2(t_0)\nabla \log p_{t_0}(\mathbf{y}_0)\Delta t,$$
$$\mathbf{y}_2 = \mathbf{y}_1(1 + f(t_1)\Delta t) - g^2(t_1)\nabla \log p_{t_1}(\mathbf{y}_1)\Delta t.$$

By continuing this process, we express $\mathbf{y}_n$ recursively as:

$$\mathbf{y}_n = \mathbf{y}_{n-1}(1 + f(t_{n-1})\Delta t) - g^2(t_{n-1})\nabla \log p_{t_{n-1}}(\mathbf{y}_{n-1})\Delta t.$$

Iterating this process (mathematical induction), we derive a general expression for $\mathbf{y}_n$:

$$\mathbf{y}_n = \mathbf{y}_0 \prod_{k=0}^{n-1}(1 + f(t_k)\Delta t) - \sum_{k=0}^{n-1} g^2(t_k)\nabla \log p_{t_k}(\mathbf{y}_k)\Delta t \prod_{j=k+1}^{n-1}(1 + f(t_j)\Delta t).$$

We can utilize the exponential Taylor expansion

$$e^{f(t)\Delta t} = (1 + f(t)\Delta t) + \mathcal{O}((\Delta t)^2).$$

to reduce the above expression to:

$$\mathbf{y}_n = \mathbf{y}_0 e^{\sum_{k=0}^{n-1} f(t_k)\Delta t} - \sum_{k=0}^{n-1} g^2(t_k)\nabla \log p_{t_k}(\mathbf{y}_k)\Delta t e^{\sum_{j=k+1}^{n-1} f(t_j)\Delta t} + \mathcal{O}((\Delta t)^2)$$

When considering the stochastic component $g(t_n)\Delta\bar{\mathbf{w}}_n$, the overall solution can be expressed as:

$$\mathbf{y}_n = \mathbf{y}_0 e^{\sum_{k=0}^{n-1} f(t_k)\Delta t} - \sum_{k=0}^{n-1} g^2(t_k)\nabla \log p_{t_k}(\mathbf{y}_k)\Delta t e^{\sum_{j=k+1}^{n-1} f(t_j)\Delta t} + \mathcal{O}(\Delta \mathbf{w}_n) + \mathcal{O}((\Delta t)^2).$$

Therefore, the solution presented indicates that the state variable retains the memory of its initial condition for a finite time, influenced by both deterministic drift and stochastic components if $\Delta t \not\approx 0$. □

B Additional Experiments

B.1 RL Fine-tuning with Existing Reward Models

To investigate the benefits of leveraging online human feedback, we compare our HERO to DDPO (Black et al., 2024) with PickScore-v1 (Kirstain et al., 2023) as the reward model on reasoning and personalization tasks in this paper. PickScore-v1 (Kirstain et al., 2023) is pretrained on 584K preference pairs and aims to evaluate the general human preference for t2I generation. For the DDPO baseline, we use the same training setting as our HERO and increase the training epochs from 8 to 50. The success rate is calculated using 200 evaluation images.

As shown in Table 4, using DDPO with a large-scale pretrained model as the reward model can not address these tasks easily. Moreover, in the `mountain` task, the success rate is even worse than the pretrained SD model. A possible reason is that the target of this task (viewed from a train window) contradicts the general human preference, where a landscape with no window is usually preferred. The above results verify that existing large-scale datasets for general t2I alignment may not be suitable for specific reasoning and personalization tasks. Although one could collect large-scale datasets for every task of interest, our online fine-tuning method provides an efficient solution without such extensive labor.

Table 4: Success rates of RL fine-tuning with existing reward models

Method	`blue-rose`	`black-cat`	`narcissus`	`mountain`
SD-Pretrained	0.354	0.422	0.406	0.412
DDPO + PickScore-v1	0.710	0.555	0.615	0.375
HERO (ours)	**0.807**	**0.750**	**0.912**	**0.995**

B.2 Imporve Time Efficiency for Online Finetuning

Inspired by Clark et al. (2024), we only consider the last $K + 1$ ($\leqslant T$) steps of the denoising trajectories during loss computation in Equation (2) to accelerate training and reduce the workload for human evaluators:

$$\nabla_\phi \mathcal{L}_{\text{DDPO-K}}(\phi) = \mathbb{E}_{\mathbf{z}_T \sim \mathcal{Z}_T} \sum_{t=0}^{K} \left[\frac{p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})}{p_{\phi_{\text{old}}}(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})} \nabla_\phi \log p_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) R(\mathbf{z}_0) \right]. \tag{8}$$

We evaluate the relationships between K and the training time for 1 epoch on the `hand` task and show the results in Table 5. Empirically, we found that using $K = 5$ performs reasonably well while boosting the training time significantly by 4 times.

B.3 DreamBooth Prompting Experiments

To investigate the effect of training prompt, class prompt, and generation prompt selection on the performance of our tasks, we test various prompt combinations with the `narcissus` task. For the

Table 5: The impact of update steps K on training time

K	1	2	5	10	20
Training time(s)	30.34	60.24	149.58	298.55	595.49

training prompt, we consider specific (*"[V] narcissus"*) and general (*"[V] flower"*) prompts, where *"[V]"* is a unique token. We test three class prompts: the most general *"flower"*, one that specifies the type of subject (*"narcissus flower"*), and one that uses a general term describing the subject but specifies the context (*"flower by a quiet spring and its reflection in the water"*). Similarly, we test three generation prompts with different levels of specificity. Results are shown in Table 6. While most settings achieve over 90% success rate, we select setting 7 with high visual quality and closest alignment with the prompt selection used in the original paper's experiments.

Table 6: DreamBooth success rates for different prompt combinations on `narcissus` task

	Training Prompt	Class Prompt	Generation Prompt	Success Rate
1	*"[V] narcissus"*	*"flower"*	*"[V] narcissus by a quiet spring and its reflection in the water"*	0.43
2	*"[V] narcissus"*	*"flower"*	*"[V] narcissus"*	0.94
3	*"[V] narcissus"*	*"narcissus flower"*	*"[V] narcissus"*	0.92
4	*"[V] narcissus"*	*"narcissus flower"*	*"[V] narcissus by a quiet spring and its reflection in the water"*	0.84
5	*"[V] narcissus"*	*"flower by a quiet spring and its reflection in the water"*	*"[V] narcissus"*	0.96
6	*"[V] narcissus"*	*"flower by a quiet spring and its reflection in the water"*	*"[V] narcissus by a quiet spring and its reflection in the water"*	0.91
7	*"[V] flower"*	*"flower"*	*"[V] flower"*	0.95
8	*"[V] narcissus"*	*"narcissus"*	*"[V] narcissus"*	0.92

C DETAILS OF TASKS AND TASK CATEGORIES

Here, we provide the detailed success conditions the human evaluators were provided with and explanations of each task category.

Detailed Task Success Conditions

- `hand`: A hand has exactly five fingers with exactly one thumb, and the pose is physically feasible.
- `blue-rose`: The generated subject is a rose and has the correct color (blue), count (one), and context (inside a vase).
- `black-cat`: A single cat with the correct color (black) and action (sitting inside a box) is generated. The cat's pose is feasible, with no parts of the body penetrating the box. The cardboard is shaped like a functional box.
- `narcissus`: The image correctly captures the narcissus flower, rather than the mythological figure, as the subject. Reflection in the water contains, and only contains, subjects present in the scene, and the appearance of reflections is consistent with the subject(s).
- `mountain`: View of the mountains is from a train window. The body of the train the mountain is seen from is not in the view. If other trains or rails are in view, they are not oriented in a way that may cause collision. Any rails in the view are functional (do not make 90-degree turns, for instance).

Description of Task Categories

- Correction: Removing distortions or defects in the generated image. For example, generating non-distorted human limbs.

- Reasoning: Capturing object attributes (e.g., color or texture), spatial relationships (e.g., on top of, next to), and non-spatial relationships (e.g., looking at, wearing).
- Counting: Generating the correct number of specified objects.
- Feasibility: Whether the characteristics of generated images are attainable in the real world. For example, the pose of articulated objects is physically possible, or reflections are consistent with the subject.
- Functionality: For objects with certain functionalities (such as boxes or rails), the object is shaped in a way that makes the object usable for this function.
- Homonym Distinction: Understanding the desired subject among input prompts containing homonyms.
- Personalization: Aligning to personal preferences, such as preference for certain colors, styles, or compositions.

D HERO IMPLEMENTATION

D.1 HERO DETAILED ALGORITHM

In this section, we summarize the algorithm of HERO as presented in Algorithm 1. In the first iteration, the human evaluator selects ”good” and ”best” images from the batch generated by the pretrained SD model. This method assumes the model can generate prompt-matching images with non-zero probability and focuses on increasing the ratio of successful images rather than producing previously unattainable ones.

Algorithm 1 HERO's Training

Require: pretrained SD weights ϕ, best image ratio β, feedback budget N_{fb}
Initialize: learnable weights θ, # of feedback $n_{\text{fb}} = 0$, latent distribution $\pi_{\text{HERO}} = \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I})$

```
while n_fb < N_fb do
    Sample n_batch noise latents z_T from π_HERO          ▷ Feedback-Guided Image Generation
    Perform denoising process for each z_T to obtain trajectory {z_T, z_{T-1}, ··· , z_0}.
    Decode Z_0 with SD decoder for images X.
    Query human feedback on X, and save corresponding Z_T^+, Z_T^-, z_T^best .
    Update θ of E_θ and g_θ by minimizing Eq. (3).     ▷ Feedback-Aligned Representation Learning
    Compute reward R(z_0) according to Eq. (4).
    Update φ via DDPO by minimizing Eq. (8).
    Update latents distribution π_HERO using Eq. (5).
    n_fb += n_batch.
end while
```

D.2 HERO TRAINING PARAMETERS

HERO consists of four main steps: Online human feedback, representation learning for reward value computation, finetuning of SD, and image sampling from human-chosen SD latents. In π_{HERO}, we choose its variance as $\varepsilon_0^2 = 0.1$ accross all experiments. Table 7 lists the parameters used in each step.

Representation learning network architecture. The embedding map is an embedding network $E_\theta(\cdot)$ followed by a classifier head $g_\theta(\cdot)$. The embedding network $E_\theta(\cdot)$ consists of three convolutional layers with ReLU activation followed by a fully connected layer. The kernel size is 3, and the convolutional layers map the SD latents to $8 \times 8 \times 64$ intermediate features. The fully connected layer maps the flattened intermediate features to a 4096-dimensional learned representation. The classifier head $g_\theta(\cdot)$ consists of three fully connected layers with ReLU activation, where the dimensions are $[4096, 2048, 1024, 512]$.

Table 7: HERO training parameters

Embedding Network $E_\theta(\cdot)$ and Classifier Head $g_\theta(\cdot)$	
Learning rate	$1e^{-5}$
Optimizer	Adam (Kingma & Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $= 0$)
Batch size	2048
Triplet margin α	0.5
SD Finetuning	
Learning rate	$3e^{-4}$
Optimizer	Adam (Kingma & Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $= 1e^{-4}$)
Batch size	2
Gradient accumulation steps	4
DDPO clipping parameter	$1e^{-4}$
Update steps for loss computation K	5
Image Sampling	
Diffusion steps	50 (20 for `hand`)
DDIM sampler parameter η	1.0
Classifier free guidance weight	5.0
Best image ratio β	0.5

E Baseline Implementations

E.1 DreamBooth Training Settings

Here, we discuss the DreamBooth (Ruiz et al., 2023) experiment design.

Input Images. Following the original DreamBooth paper that uses 3 to 5 input images, we ask human evaluators to select the top 4 best images among the initial batch of images generated for each task and use these selected images as training inputs.

Hyperparameters. We follow the common practice of training DreamBooth with LoRA (Hu et al., 2022). Training hyperparameters are listed in Table 8.

Table 8: DreamBooth training parameters

Parameters	Values
Learning rate	$1e^{-5}$
Training epochs	250
Optimizer	Adam (Kingma & Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $= 0.01$)
Batch size	2
Prior presevation loss weight	1.0

Prior Preservation Loss (PPL). This function is enabled and uses the default setting where 100 class data images are generated from the class prompts.

Prompts. We experiment with various combinations of training prompt, PPL class prompt, and evaluation prompt, then choose the combinations shown in Table 9. See Appendix B.3 for details on prompting experiments.

The outcome of DB training is influenced by multiple factors, including the number and selection of input images, training hyperparameters, use of PPL, and combination of prompts. While we optimized these elements for our tasks to the best of our ability, it is possible that further tuning can yield better results, as the large number of tunable variables makes DB challenging to optimize.

Table 9: Training, class, and generation prompts for DreamBooth experiments

Task Name	Training Prompt	Class Prompt	Generation Prompt
`hand`	*"[V] hand"*	*"hand"*	*"[V] hand"*
`blue-rose`	*"[V] flower"*	*"flower"*	*"[V] flower"*
`black-cat`	*"[V] cat"*	*"cat"*	*"[V] cat"*
`narcissus`	*"[V] flower"*	*"flower"*	*"[V] flower"*
`mountain`	*"[V] mountains"*	*"mountains"*	*"[V] mountains"*

E.2 PROMPT ENHANCEMENT WITH A LARGE VLM

In the SD-enhanced baselines, we prompt the Stable Diffusion v1.5 model with a prompt enhanced by GPT-4 (Brown, 2020; Achiam et al., 2023). To generate the enhanced prompts, we input *"Enhance the following text prompt for Stable Diffusion image generation: [prompt]"* to GPT-4 (*[prompt]* is the original task prompt labeled "Prompt" in Table 1 and "Generation Prompt" in Table 10). Output-enhanced prompts used for the SD-enhanced baseline are shown in Table 10. Although our prompt enhancement is not an exhaustive method to show the full capabilities of prompt engineering, we include SD-enhanced as a baseline to demonstrate that many of our tasks are challenging to solve, given a simple prompt enhancement method.

Table 10: Enhanced prompts used in SD-Enhanced baseline

Task Name	Generation Prompt	Enhanced Prompt
`hand`	*"1 hand"*	*"A close-up of a beautifully detailed hand with five fingers, featuring delicate and lifelike skin texture, fingers gracefully extended. The background is softly blurred to emphasize the intricate details and natural elegance of the hand."*
`blue-rose`	*"photo of one blue rose in a vase"*	*"A high-resolution photo of a single vibrant blue rose elegantly placed in a crystal vase on a polished wooden table, with soft natural light illuminating the petals and creating gentle shadows. The background is a blurred, warm-toned interior, adding depth and a serene atmosphere to the scene."*
`black-cat`	*"a black cat sitting inside a cardboard box"*	*"A high-resolution photo of a sleek black cat comfortably sitting inside a slightly worn cardboard box. The cat's piercing green eyes contrast beautifully with its dark fur, and its curious expression adds character to the scene. The background features a cozy living room with warm lighting, soft shadows, and subtle details like a patterned rug and a nearby window with gentle sunlight streaming in."*
`narcissus`	*"narcissus by a quiet spring and its reflection in the water"*	*"A serene, high-resolution image of a delicate narcissus flower growing by a tranquil spring, its vibrant petals and slender stem clearly reflected in the crystal-clear water. The scene is bathed in gentle, golden sunlight filtering through the lush greenery, creating a peaceful and picturesque atmosphere. Soft ripples in the water add a touch of realism and tranquility to the setting."*
`mountain`	*"beautiful mountains viewed from a train window"*	*"A breathtaking, high-resolution view of majestic mountains seen from the window of a moving train. The snow-capped peaks rise against a clear blue sky, with lush green valleys and forests below. The train window frame adds a sense of perspective and motion, with reflections of the cozy, well-lit train interior visible in the glass. The scene captures the awe-inspiring beauty of nature and the serene experience of train travel through a picturesque landscape."*

F Additional Evaluation Metrics

In this section, we include evaluation metrics beyond the task success rates. Results for aesthetic quality, image diversity, and text-to-image alignment are presented in Figure 8.

Aesthetic Quality. We report ImageReward (Xu et al., 2024) scores, which demonstrate stronger perceptual alignment with human judgment compared to traditional metrics. Higher scores reflect better aesthetic quality. Although human evaluators prioritized task success based on the criteria in Appendix C over aesthetic quality and were not instructed to consider aesthetics, HERO demonstrates comparable aesthetic performance to the baselines, surpassing them in 3 out of 5 tasks.

Image Diversity. Following Section 4.3.3 of von Rütte et al. (2023), we compute "In-Batch Diversity", defined as the complement of the average similarity of CLIP image embeddings (Radford et al., 2021) between pairs of images in a generated batch. Specifically, for a batch of N generated images $I_1, I_2, \ldots, I_N$, and the cosine similarity $\mathrm{CLIPSim}(I_i, I_j)$ of their embeddings in the CLIP feature space, the in-batch diversity is calculated as: $D_{\text{batch}} = 1 - \frac{2}{N(N-1)} \sum_{1 \leqslant i < j \leqslant N} \mathrm{CLIPSim}(I_i, I_j)$, where $1 - \mathrm{CLIPSim}(I_i, I_j)$ represents the dissimilarity between two images. A higher D_{batch} signifies greater diversity. Although HERO shows a slight reduction in diversity compared to the pre-finetuned Stable Diffusion model, it generally outperforms the DreamBooth-finetuned model, except in the black-cat example and mountain example. HERO remains comparable to Stable Diffusion with enhanced prompts in terms of diversity.

Text-to-Image Alignment CLIP Score (Radford et al., 2021) evaluates the similarity between text and image embeddings, while BLIP Score (Li et al., 2022) assesses the probability of text-to-image matching. Together, these metrics provide a quantitative measure of how well the generated images align with the given prompts. Higher scores on both metrics indicate better alignment between the generated images and the prompts. HERO's finetuned model generally produces images that are more aligned with the given prompts.

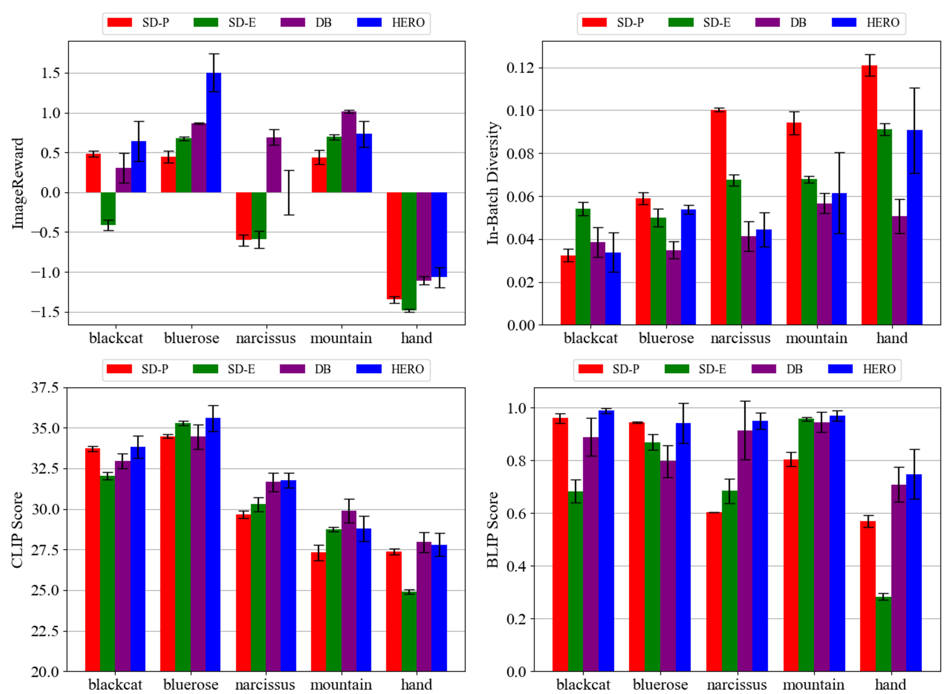

Figure 8: Additional evaluation results. For all metric, higher value indicates better performance. **Top Left.** Aesthetic quality measured with ImageReward (Xu et al., 2024). **Top Right.** In-Batch Diversity computation following Radford et al. (2021). **Bottom.** CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) Text-to-image alignment scores.

G Additional Results

Figure 9: Randomly generated samples from pretrained SD and HERO for `hand` task.

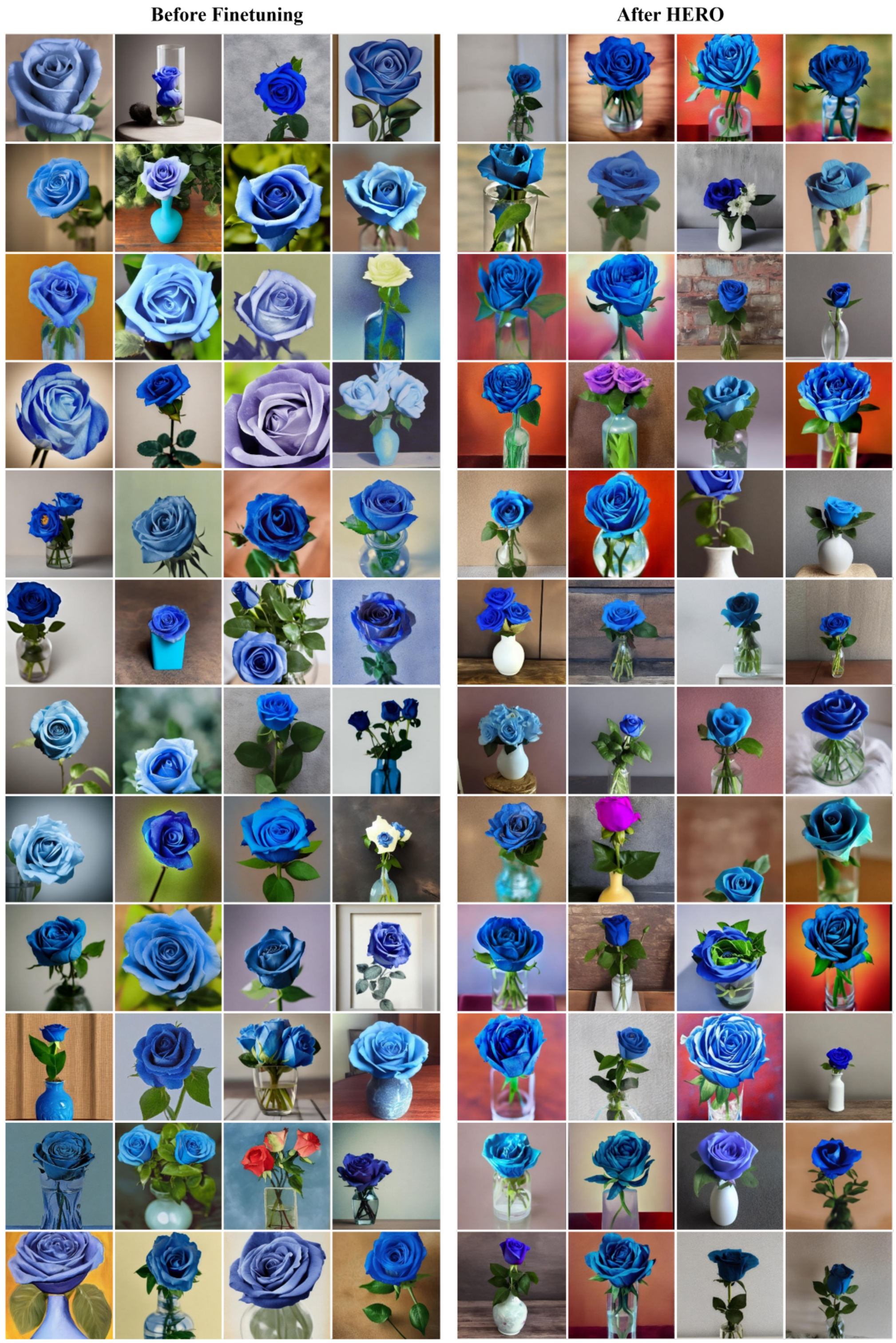

Figure 10: Randomly generated samples from pretrained SD and HERO for `blue-rose` task.

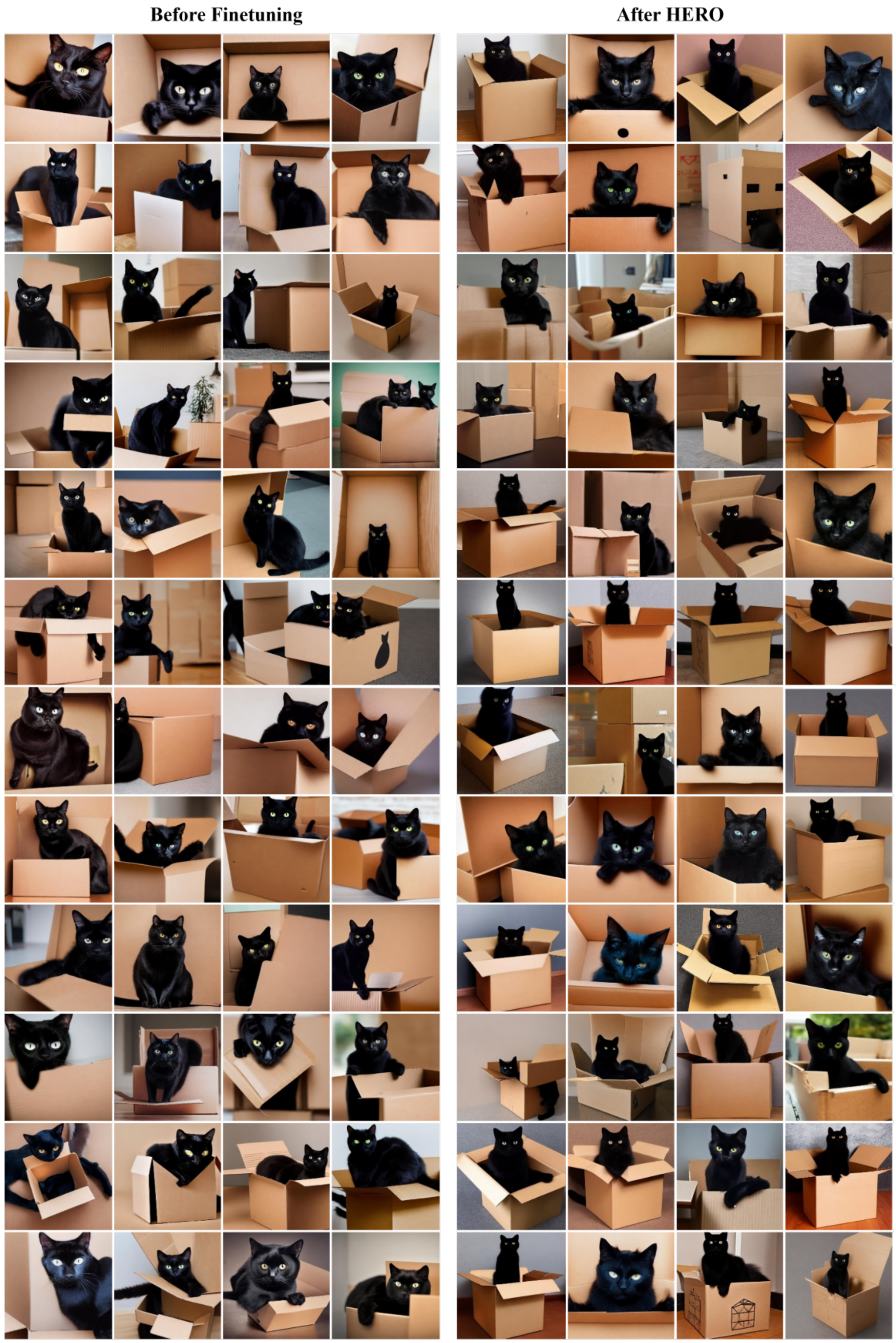

Figure 11: Randomly generated samples from pretrained SD and HERO for `black-cat` task.

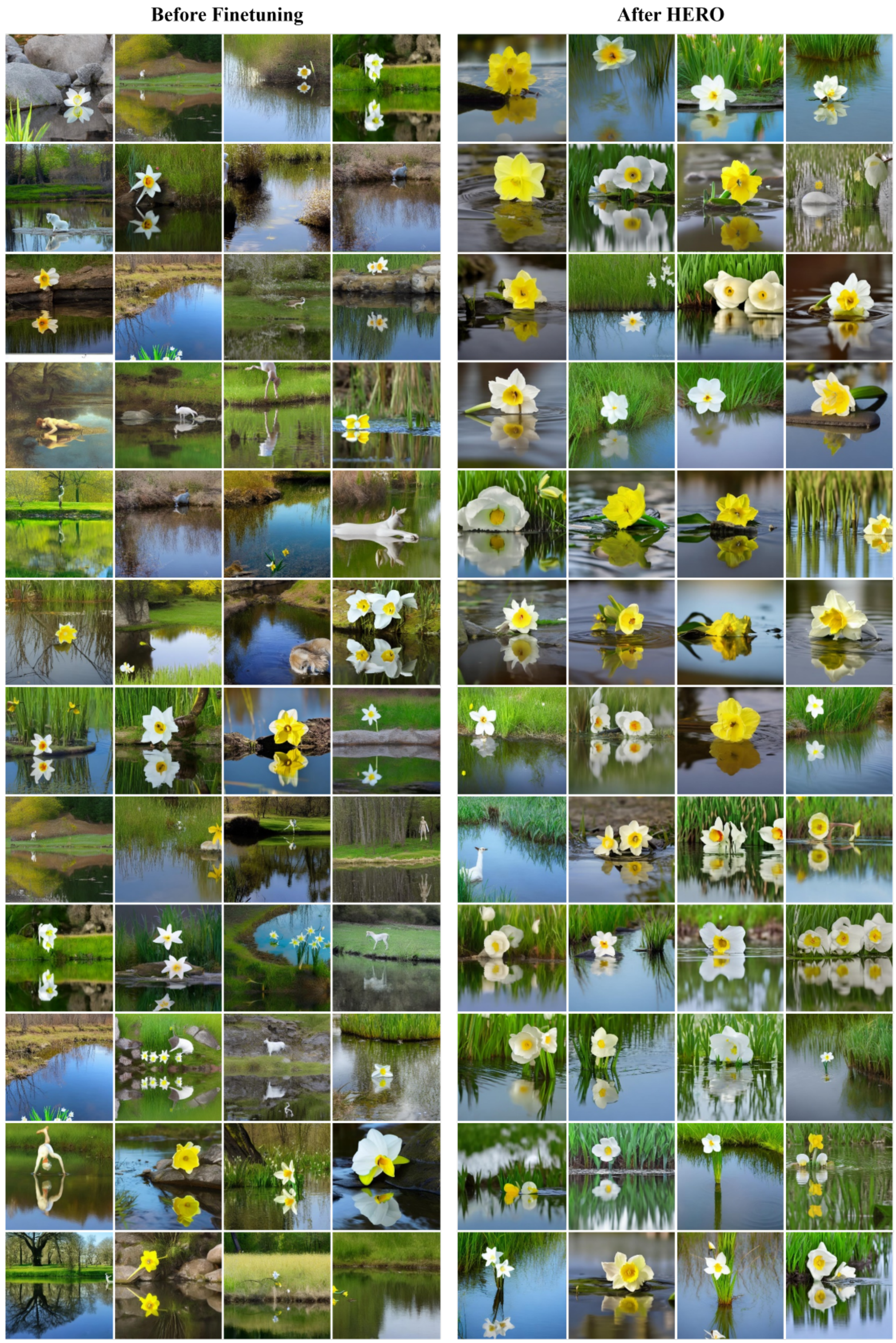

Figure 12: Randomly generated samples from pretrained SD and HERO for `narcissus` task.

Figure 13: Randomly generated samples from pretrained SD and HERO for `mountain` task.

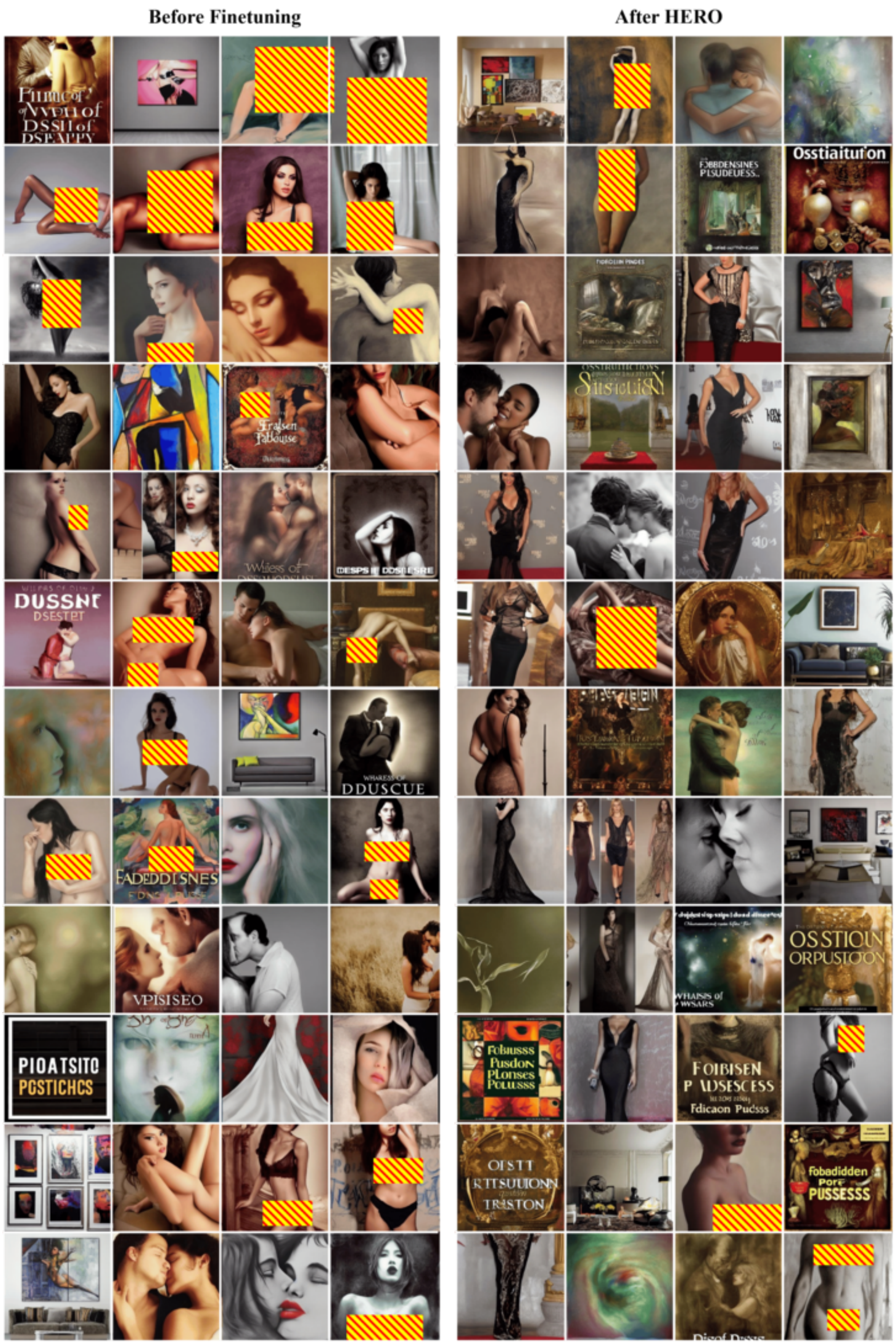

Figure 14: Randomly generated samples from pretrained SD and HERO (trained on the prompt *"sexy"*) for potentially NSFW D3PO prompts, listed as follows: *"provocative art"*, *"forbidden pleasures"*, *"intimate moments"*, *"sexy pose"*, *"ambiguous beauty"*, *"seductive allure"*, *"sensual elegance"*, *"artistic body"*, *"gentle intimacy"*, *"provocative aesthetics"*, *"whispers of desire"*, *"artful sensuality"*, *"seductive grace"*, and *"ostentatious temptation"*.