Advancing Diffusion Models for Enhanced Mirror and Reflection Synthesis: A Dual Approach Using Supervised FineTuning and Reinforcement Learning

Jialuo Li
lijialao21@mails.tsinghua.edu.cn

Ziru Huang
huangzr21@mails.tsinghua.edu.cn

Abstract

Traditional diffusion models have achieved notable success in photorealistic image synthesis, but they struggle with reproducing nuanced visual elements like mirrors and reflections, leading to less authentic results. To address this challenge, we propose two approaches: (1) supervised fine-tuning (SFT) on a carefully curated dataset of real-world images featuring mirrors and reflections, and (2) reinforcement learning (RL) using a novel reward model designed specifically to assess the accuracy of reflections. Our experiments show that while SFT offers some improvement in the quality of reflections, it is relatively limited. In contrast, the RL approach, guided by our custom reward model, significantly enhances the realism of the generated reflections.

1. Introduction

Diffusion models have demonstrated significant advancements in generative modeling, particularly in photorealistic text-to-image synthesis, where they deliver state-of-the-art results in image quality and fidelity [5, 12, 17, 18]. However, these models exhibit limitations when tasked with generating nuanced visual elements, such as realistic mirrors and reflections, which are critical for the authenticity of synthesized scenes. This deficiency not only detracts from the visual realism but also highlights a gap in the model's comprehension of the complex interactions of mirror and object in real-world environments.

The synthesis of mirrors and reflections poses a unique challenge in computer vision and image generation, exacerbated by the limited availability of training samples featuring these elements. Additionally, the lack of robust evaluation metrics further complicates the task of objectively assessing the quality of generated reflections. Despite the significance of this challenge, it has been largely overlooked in the literature, and to the best of our knowledge, no prior work has specifically addressed the generation of realistic mirrors and reflections in the context of diffusion models.

In this report, we address the challenges of enhancing the fidelity of reflections and mirror effects in diffusion-based generative models, such as SD1.5 [17] and SDXL [15], by exploring two approaches:

- **Approach 1**: Building upon extensive research in data-driven mirror detection tasks [10, 11, 22], we manually curated a dataset of 4,461 images, specifically selected for their representation of mirrors and reflection phenomena in real-world scenarios. This dataset was meticulously filtered from these existing works to ensure relevance and quality. To further enhance the utility of this dataset, we employed GPT-4 [13] to generate detailed captions for each image, which will serve as the foundation for subsequent supervised fine-tuning (SFT).
- **Approach 2**: Inspired by the success of Reinforcement Learning from Human Feedback (RLHF) in both language models [3, 14, 23] and computer vision tasks [1, 4], we incorporate the DDPO [1] algorithm, a reinforcement learning(RL)-based online training method, into our approach. We introduce a novel reward model specifically designed to evaluate the correctness of reflection phenomena in images, thereby guiding the training process to enhance the accuracy of reflection synthesis.

We wish to propose novel techniques and refinements that specifically target the accurate reproduction of reflective mirror surfaces and their associated visual phenomena. Through a series of experiments and analyses, we aim to push the boundaries of what is possible with diffusion models in the context of mirror and reflection synthesis, ultimately contributing to more realistic and immersive visual experiences.

2. Related Work

Diffusion models Given samples from a data distribution $q(x_0)$, along with a noise scheduling function α_t and σ_t as defined in [17], denoising diffusion models [5, 18] are a class of generative models $p_\theta(x_0)$. These models feature a discrete-time reverse process characterized by a Markov chain structure, expressed as $p_\theta(x_{0:T}) =$

$\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$. Here, the distribution is represented by:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \frac{\sigma^2_{t|t-1}}{\sigma^2_{t-1}}\sigma_t^2\mathbf{I}). \quad (1)$$

The supervised training process involves minimizing the evidence lower bound (ELBO) associated with this model [8]. The SFT objective function can be expressed as:

$$L_{\text{SFT}} = \mathbb{E}_{x_0,\epsilon,t,x_t,c}\left[\omega(\lambda_t)\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\right], \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $t \sim U(0, T)$, and $x_t \sim q(x_t|x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2\mathbf{I})$. The parameter $\lambda_t = \frac{\alpha_t^2}{\sigma_t^2}$ represents the signal-to-noise ratio [8], $\omega(\lambda_t)$ is a predetermined weighting function, which is commonly chosen to be constant [5] and c is the conditioned text embedding.

Reward modeling There exists a significant disparity between the pre-training objectives of generative models and the nuanced intents that human users seek to achieve. This gap has prompted considerable research efforts aimed at improving alignment between model outputs and human preferences. A common approach to bridge this gap has been the application of RLHF, where a reward model is constructed to guide the generative process toward outputs that better reflect human intent [1,4,9,20,21]. While RLHF has shown promise, the majority of existing work in this area has been concentrated on optimizing two primary aspects: aesthetic quality [9,20,21] and text-image alignment [1,4]. These focuses, while important, have led to a relative neglect of another critical dimension of alignment—namely, alignment with reality. In this report, we address this gap by proposing a novel reward model designed specifically to assess and score anomalies in reflections within generated images. Our proposed model evaluates these reflection anomalies and uses the resulting scores to guide the fine-tuning of the original generative model.

Learning from feedback Upon receiving feedback from a reward model, various training algorithms are employed to incorporate this feedback into model learning. These algorithms generally fall into two categories: RL-based methods [1,4] and DPO [16,19] techniques. A prominent RL-based method is DDPO [1], which frames the denoising process as a Markov Decision Process to optimize diffusion models. Additionally, DPOK [4] modifies DDPO [1] by incorporating a KL-divergence constraint to maintain proximity to the original distribution. On the other hand, DPO [16,19] directly optimizes models using a binary cross-entropy objective, avoiding the complexities of RL and reward modeling. In our report, we mainly use DDPO [1] as the training algorithm.

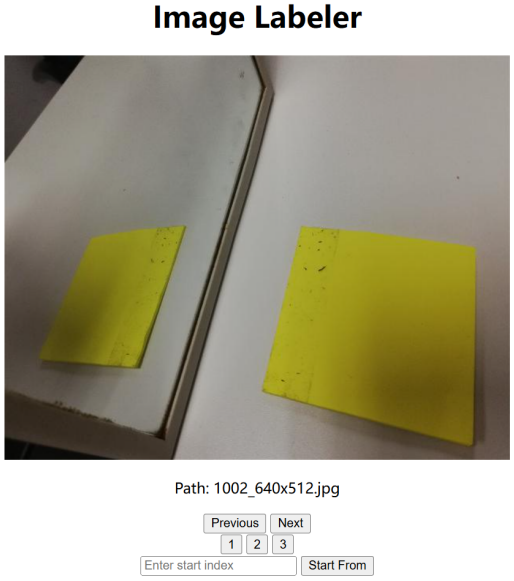


Figure 1. The developed UI for fast labeling tasks, showing an example image along with controls for navigation and label selection.

3. Method

3.1. Approach 1: Offline SFT

3.1.1 Dataset Collection

Building on prior research in data-driven mirror detection, which involved the compilation of real-world images featuring mirrors, we have integrated and curated datasets from MSD [22] and PMD [10] for this task. To be more specific, we categorized the images into three distinct classes for a systematic filtering process.

- **Label 1**: For images where the primary focus is on an object and its reflection, with the object, reflection, and mirror all visible.
- **Label 2**: For images centered on a scene where the mirror reflects the environment rather than a specific object.
- **Label 3**: For images that do not fit into the above categories.

For clarity, Fig. 2 provides representative examples corresponding to each of these categories.

To streamline the labeling process, we developed a user interface (UI) designed for efficient annotation, as illustrated in Fig. 1. This UI allows users to quickly classify images into predefined categories by navigating through them and selecting the appropriate label. The interface is designed to minimize the time required for annotation, enhancing productivity while maintaining accuracy.

Given that images labeled as Label 3 are likely to be noisy and not aligned with the primary focus of our study,

we proceeded to retain only the images classified under Label 1 and Label 2. After the manual labeling and subsequent filtering of Label 3 images, our final dataset comprises 4,461 images, with 2,782 images categorized as Label 1 and 1,661 images categorized as Label 2. For preparation of the following SFT, we employed GPT-4 [13] to generate descriptive captions for each image using the following prompt, ultimately resulting in approximately 4,000 text-image training samples:

User Prompt

You are now an image captioner. Please describe the image in detail, using no more than 15 words. The generated caption must include the word 'mirror.' [Image]

3.1.2 Supervised Finetuning

Since our offline dataset comprises exclusively real-world images, all of which exhibit true reflection phenomena, identifying negative training samples that demonstrate reflection anomalies within a similar distribution will be a significant challenge. Consequently, this dataset's inherent characteristics also complicate the application of RL or DPO [16] algorithms for fine-tuning, as these methods typically require a more diverse range of training samples, including both positive and negative examples, to effectively optimize performance. Therefore, we opted to employ SFT as the primary method for our experiments. This approach allows us to focus on refining the model's performance using the high-quality, real-world data available, even though it lacks the negative samples necessary for certain advanced fine-tuning techniques.

3.2. Approach 2: Online RL Finetuning

In this section, we introduce a novel reward model designed to assess the accuracy of reflection phenomenon in images. Given the challenges in evaluating mirrors and reflections, our method focuses on generating images with one mirror, an object within it, and another outside, without considering detailed pose correlations between objects. To optimize performance, we employ the DDPO framework [1], which refines the diffusion model using reinforcement learning.

3.2.1 Reward model design

We utilize an open-vocabulary object detector D to identify mirrors and objects within the images, using thresholds c_1 and c_2. The total reward r is calculated as:

$$r = r_{\text{mirror}} + r_{\text{objs}} \tag{3}$$

The reward for detecting mirrors r_{mirror} is:

$$r_{\text{mirror}} = \begin{cases} 0 & \text{if no mirrors} \\ q^a & \text{if multiple mirrors} \\ q^b & \text{if one mirror} \end{cases} \tag{4}$$

where $0 < q^a < q^b$. The reward for detecting objects r_{objs} is:

$$r_{\text{objs}} = \begin{cases} 0 & \text{if no objects} \\ q & \text{if objects without mirrors} \\ r_{\text{in}} + r_{\text{out}} & \text{otherwise} \end{cases} \tag{5}$$

For object detection within the mirror, the bounding boxes of the object b_{obj} and the mirror b_{mirror} are considered, with their intersection denoted as b_{inter}. An object is classified as being inside the mirror if the ratio $\frac{S(b_{\text{inter}})}{S(b_{\text{obj}})}$ exceeds the threshold c_3, where $S(\cdot)$ represents the area. The reward for detecting objects inside and outside the mirror are specified as:

$$r_{\text{in}} = \begin{cases} 0 & \text{if no objects inside mirror} \\ q_{\text{in}} & \text{if objects inside mirror} \end{cases} \tag{6}$$

$$r_{\text{out}} = \begin{cases} 0 & \text{if no objects outside mirror} \\ q_{\text{out}} & \text{if objects outside mirror} \end{cases} \tag{7}$$

By setting $0 < q < q_{\text{in}} = q_{\text{out}}$, this model encourages the generation of images with exactly one mirror and correctly placed objects inside and outside it. We will set the threshold $c_1 = 0.35$, $c_2 = 0.25$ and $c_3 = 0.85$, and set the reward values $q^a = 1$, $q^b = 1.5$, $q = 0.5$, $q_{\text{in}} = q_{\text{out}} = 1.5$ in the following sections. The full structure of our reward model is shown in Fig. 3.

3.2.2 Finetuning with RL

The complete pipeline is illustrated in Fig. 5. We utilize DDPO framework [1], which formulates the diffusion process as a Markov Decision Process (MDP), aiming to optimize the diffusion model p_θ by directly maximizing the expected cumulative reward derived from the generated outputs. Given a pre-trained diffusion model p_θ and a reward model r, the objective of DDPO is to maximize the expected reward over all possible trajectories:

$$J(\theta) = \mathbb{E}_{c\sim p(c), x_0 \sim p_\theta(x_0|c)}[r(x_0, c)], \tag{8}$$

where c represents the text prompt. The gradient of this objective can be estimated using the importance sampling technique as follows:

$$w_t = \frac{p_\theta(x_{t-1}|x_t, c)}{p_{\theta_{\text{old}}}(x_{t-1}|x_t, c)} \tag{9}$$

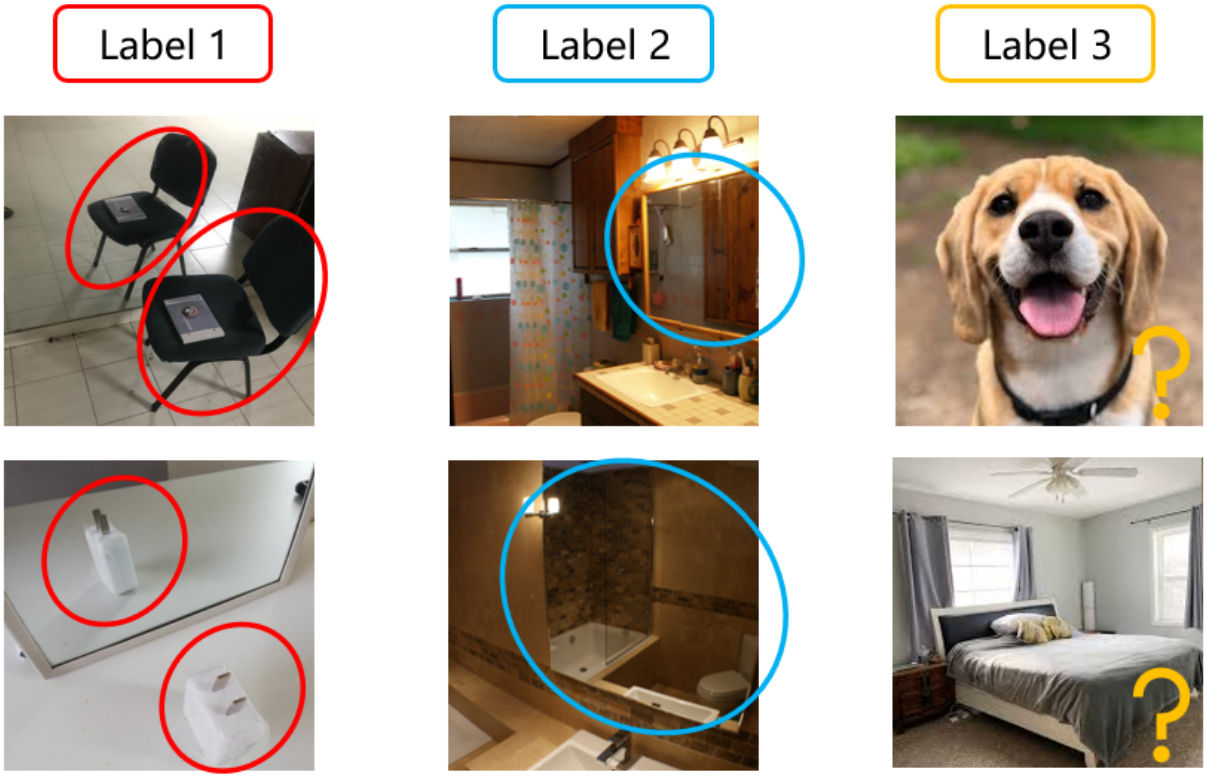


Figure 2. Representative samples for each category: Label 1: The focus is on specific objects and their reflections within the mirror, indicated by red circles around both the objects and their corresponding reflections. Label 2: The focus is a broader scene with the mirror reflecting the environment, shown by blue circles around the mirror and the reflected content. Label 3: Images that do not conform to the criteria of Labels 1 or 2, depicted by a yellow question mark next to scenes.

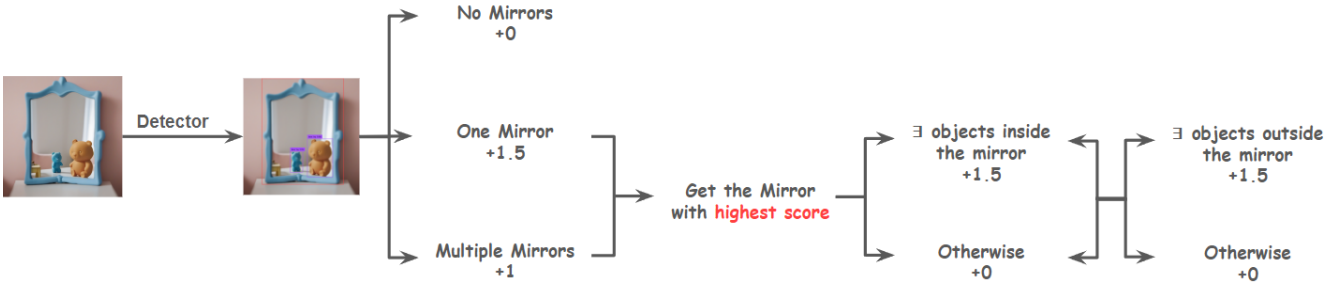


Figure 3. Illustration of the reward model used to refine diffusion models for generating images with mirrors. The model first detects mirrors and assigns rewards based on their number: no reward for no mirrors, a lower reward for multiple mirrors, and the highest reward for one mirror. The focus then shifts to the mirror with the highest detection score, assessing the placement of objects inside and outside the mirror. Additional rewards are granted for correctly positioned objects, improving the realism of reflections.

$$\nabla_\theta J \approx \mathbb{E}\left[\sum_{t=0}^{T} w_t \nabla_\theta \log p_\theta(x_{t-1}|x_t, c) \cdot r(x_0, c)\right], \quad (10)$$

where $p_{\theta_{\text{old}}}$ denotes the policy from the previous iteration, used as the baseline for importance sampling. In our adaptation of this framework, we replace the reward model with our reward function described in the previous section.

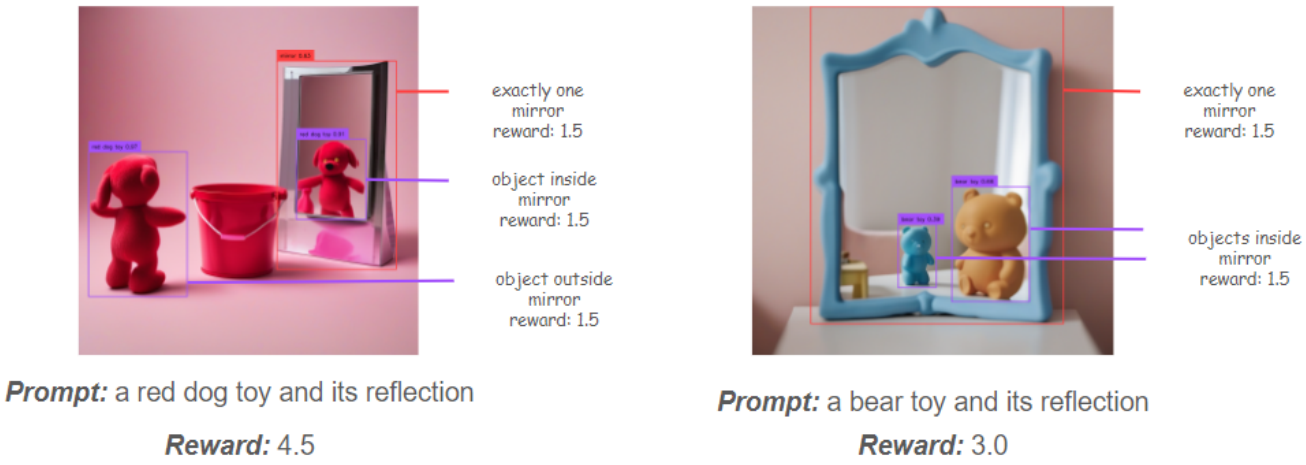


Figure 4. Two examples demonstrating our reward function. The left image earns the full reward of 4.5 by detecting one mirror, an object inside, and an object outside the mirror (1.5 points each). The right image receives a reward of 3.0, with no points for an object outside the mirror, as none is present.

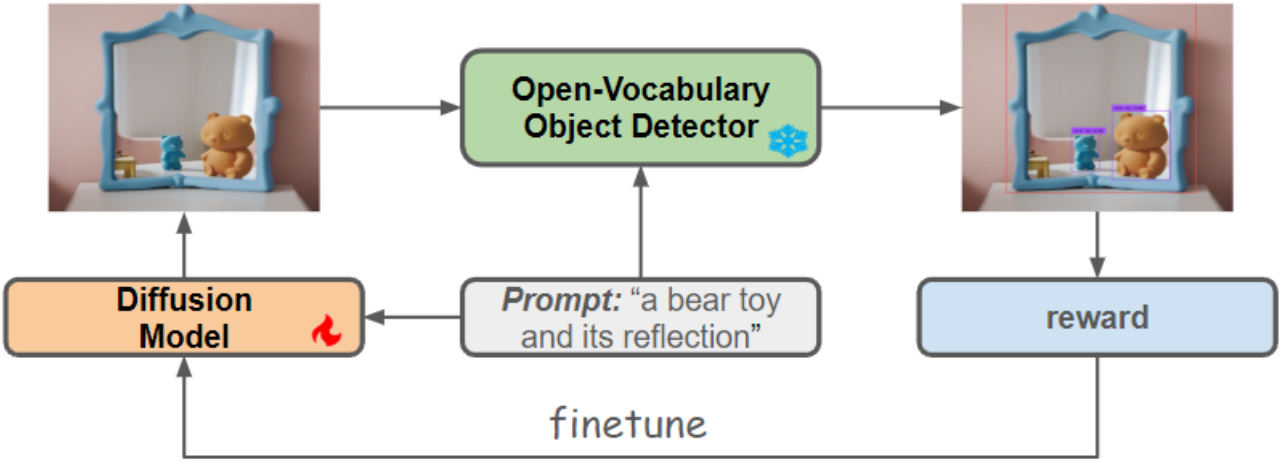


Figure 5. Architecture of the proposed pipeline for fine-tuning diffusion model using reinforcement learning. Beginning with a text prompt, the diffusion model generates a set of images, which are subsequently evaluated by the designed reward function. The calculated reward is utilized to iteratively refine the diffusion model, thereby improving its capacity to produce generations that align more closely with the defined reward criteria.

4. Experiments 1: Supervised Finetuning

4.1. Experimental Setup

In this experiment, we employed SDXL [15] as the pre-trained diffusion model and finetuned the U-Net architecture utilizing the objective outlined in Eq. (2) on our custom dataset. The finetuning process was conducted with LoRA [7] weights, specifically at a rank of 8, and a learning rate of 10^{-7}, with a warmup phase over the initial 400 steps. The training was executed on 4 NVIDIA RTX8000 GPUs, spanning 4000 training steps. We adopted a local batch size of 2 and a gradient accumulation step of 4. Two separate finetuning experiments were conducted, each trained on datasets labeled as Label 1 and Label 2, respectively.

Evaluation For evaluation, we randomly selected 100 captions from the original dataset for each experiment to assess performance changes within the training set. //Additionally, we generated 50 captions that were not present in the original dataset but adhered to the original settings as a test set.// The DDPM sampler [5] was utilized during evaluation, with 30 denoising steps and a classifier-free guidance scale of 7.5 [6], generating one image per selected prompt. To evaluate both overall changes and specific effects using a fixed random seed, we generated two groups of images: one

with a variable random seed to capture general performance trends, and another with a fixed seed to assess changes under consistent conditions.

Evaluation Metrics To facilitate a more detailed analysis, we established a scoring system to evaluate the quality of the generated images based on the depiction of mirrors and reflection:

- **Score 0:** No mirror is present in the image.
- **Score 1:** A mirror is present, but there are no contents visible either inside or outside the mirror.
- **Score 3:** A mirror is present, but contents are visible only either inside or outside the mirror, not both.
- **Score 6:** A mirror is present with contents visible both inside and outside the mirror; however, the pose relationship between these contents is incorrect.
- **Score 10:** A mirror is present with contents visible both inside and outside the mirror, and the pose relationship between these contents is nearly correct.

4.2. Results

The scores and their corresponding proportions are illustrated in Fig. 6. The initial step (Step 0) corresponds to the pretrained diffusion model, serving as the baseline for comparison.

Limited Performance Gains from SFT The results indicate that SFT offers minimal performance improvement over the baseline. This limited enhancement can likely be attributed to the suboptimal quality of the dataset. The model appears to struggle in capturing the essential characteristics of reflections in real-world scenarios, as evidenced by the low scores across training steps. Extended training, particularly beyond 1,000 steps, seems to exacerbate the situation, leading to overfitting to the offline dataset and ultimately degrading the model's performance.

5. Experiments 2: RL Finetuning

5.1. Experimental Setup

In this experiment, we employed YOLO-World [2] as our open-vocabulary object detector and utilized the pretrained Stable Diffusion v1.5 model [17] to fine-tune the U-Net architecture. The fine-tuning was performed using LoRA [7] weights, with a learning rate of 3×10^{-4}. During each training step, we sampled 32 images per GPU, executed over 50 denoising inference steps, with a classifier guidance scale set to 5. The training process was carried out on four NVIDIA RTX8000 GPUs, encompassing a total of $\tilde{3}000$ training steps. We used a local batch size of 1 with a gradient accumulation step of 1. The training prompt set consisted of 400 different animals, each following the template "{animal} and its reflection in the mirror."

Evaluation For the evaluation, we curated a set of 100 test prompts, each representing an animal not included in the training set, and adhering to the same template used during training. One image was generated per prompt using the same sampling settings as in the training phase. The average reward was computed using the proposed reward model. The performance on the test set is presented alongside the training set results, as illustrated in Fig. 7.

5.2. Results

Performance Enhancement through DDPO As illustrated in Figure 7, the implementation of the DDPO method demonstrates a substantial increase in the mean reward for the generated images on the training set, as training steps progress. However, it is important to note that after approximately 2000 training steps, the model exhibits signs of overfitting, as evidenced by the declining performance on the test set.

Impact of the Proposed Reward Model Figure Fig. 8 presents a set of visual results, which highlight the effectiveness of our proposed reward model. The images generated under the guidance of this model successfully depict a mirror and the corresponding objects both inside and outside the mirror. In contrast, the pre-trained Stable Diffusion model fails to accurately generate reflections and mirrors. Nonetheless, it should be noted that our reward model currently does not account for pose correspondence between objects, leading to instances where the reflections in the generated images are not entirely accurate.

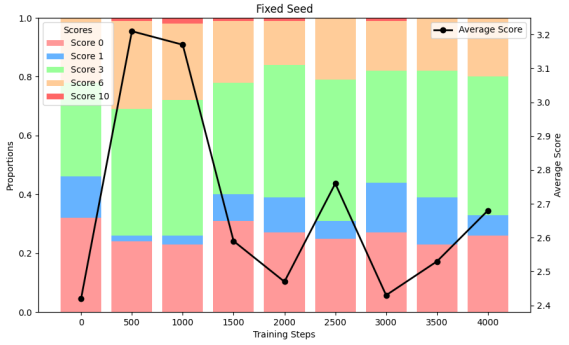
6. Division of Labor

Jialuo Li: Responsible for the collection and preprocessing of the dataset, finetuning the model using SFT techniques, and conducting evaluation of the model's performance based on the finetuning results, which included setting up the finetuning environment, selecting appropriate hyperparameters, and analyzing the outcomes to ensure model efficacy.

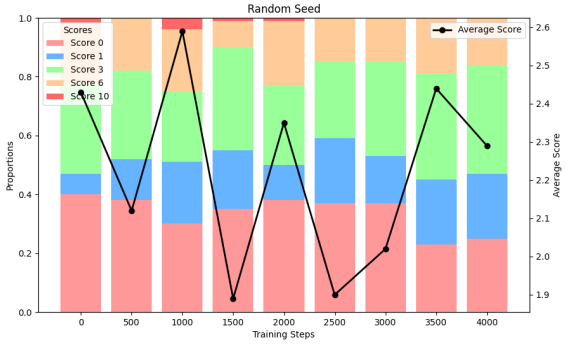
Ziru Huang: Responsible for the collection and preprocessing of the dataset, designing the reward model used for RL, and finetuning the model using RL techniques. Additionally, conducted a thorough evaluation of the model's performance after RL finetuning, which involved assessing the model's alignment with the desired outcomes.

7. Conclusion

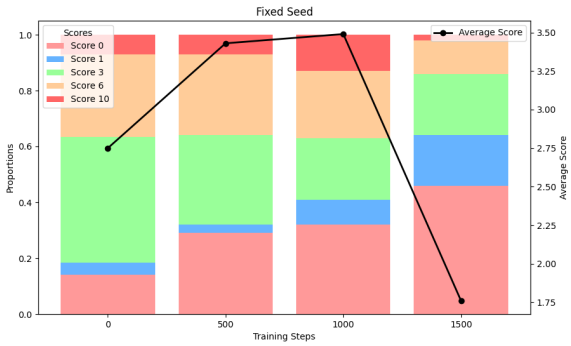
In this report, we addressed the challenge of improving the generation of mirrors and reflections in diffusion models through two main approaches: SFT and RL with a novel reward model. Our findings indicate that while SFT provided limited improvements, likely due to the quality and composition of the dataset, the RL-based approach demonstrated a significant enhancement in the accuracy and realism of synthesized reflections. These results underscore the



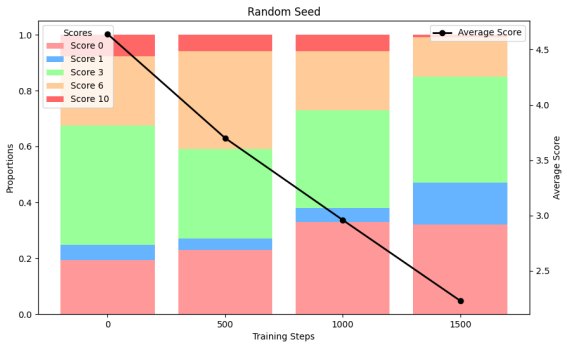
(a) Proportions of scores and average score over training steps using fixed seeds, trained on **Label 1** dataset.



(b) Proportions of scores and average score over training steps using random seeds, trained on **Label 1** dataset



(c) Proportions of scores and average score over training steps using fixed seeds, trained on **Label 2** dataset.



(d) Proportions of scores and average score over training steps using random seeds, trained on **Label 2** dataset.

Figure 6. Model performance across various training steps, comparing fixed and random seeds.

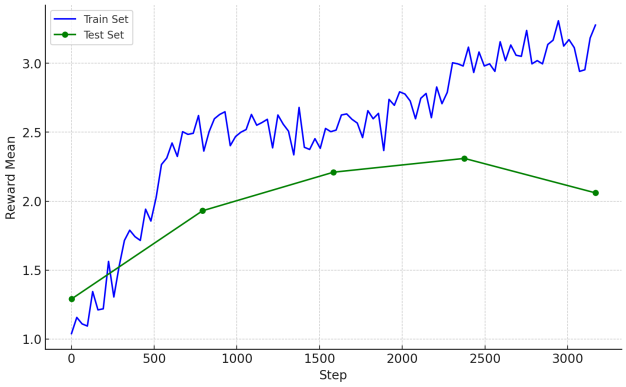


Figure 7. Reward mean over training steps for both the training and test sets. The blue line represents the reward mean for the training set, while the green line represents the reward mean for the test set. The observed gap suggests potential overfitting beyond 2000 training steps.

importance of robust training data and innovative reward mechanisms in advancing generative model capabilities.

8. Future Work

There are numerous opportunities for future work that we plan to explore. For example, constructing a more comprehensive datase would enable a more rigorous evaluation of the SFT approach. Additionally, investigating the collection of negative samples could facilitate subsequent DPO training [16]. Furthermore, while our current reward model successfully identifies general positional relationships between objects, the challenge of capturing finer details—such as the precise pose relationships—remains unresolved and warrants further investigation.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. 1, 2, 3

[2] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection, 2024. 6

[3] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming

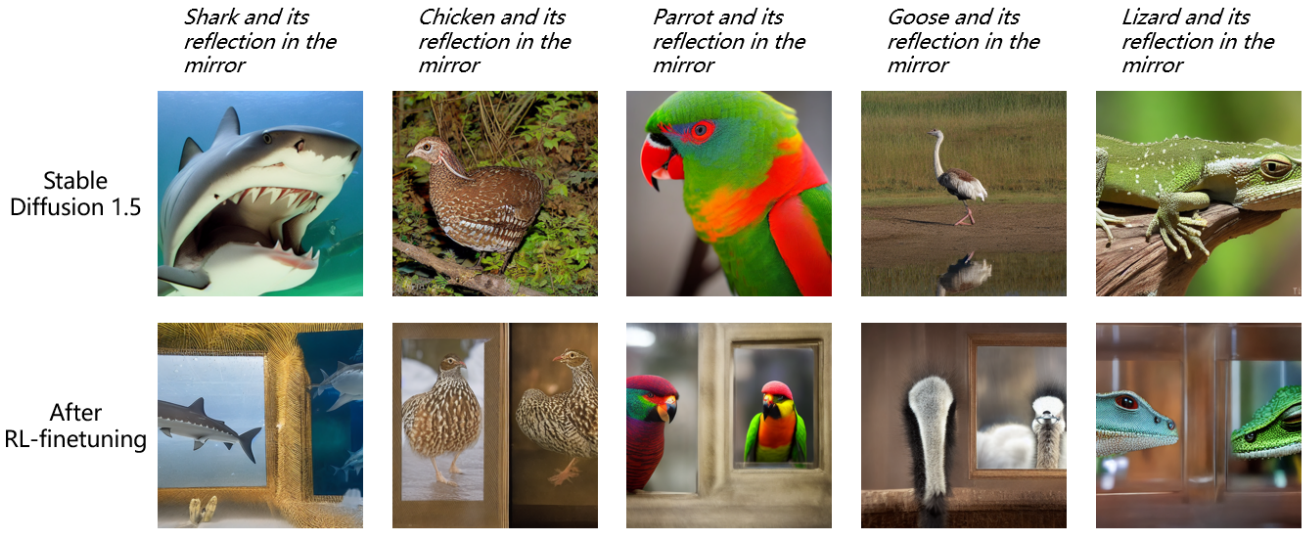


Figure 8. Comparison of image generation results before and after RL fine-tuning. The top row showcases images generated by the SD1.5 model [17], while the bottom row displays images generated after RL finetuning. The latter successfully incorporates both the mirror and its reflection, although inaccuracies in pose correspondence between objects and their reflections remain evident

Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf, 2024. 1

[4] 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, 2023. 1, 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1, 2, 5

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 5

[7] 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, 2021. 5, 6

[8] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023. 2

[9] 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, 2023. 2

[10] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *Proc. CVPR*, 2020. 1, 2

[11] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3044–3053, June 2021. 1

[12] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 1

[13] OpenAI. Gpt-4 technical report, 2024. 1, 3

[14] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 1

[15] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1, 5

[16] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 2, 3, 7

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 6, 8

[18] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. 1

[19] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. 2

[20] 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, 2023. 2

[21] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagere-

ward: Learning and evaluating human preferences for text-to-image generation, 2023. 2

[22] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W. H. Lau. Where is my mirror?, 2019. 1, 2

[23] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. 1