The Image as Its Own Reward: Reinforcement Learning with Adversarial Reward for Image Generation

Weijia Mao[1], **Hao Chen**[2†], **Zhenheng Yang**[2], **Mike Zheng Shou**[1†]

[1]Show Lab, National University of Singapore, [2]ByteDance

[†]Corresponding authors

Abstract

A reliable reward function is essential for reinforcement learning (RL) in image generation. Most current RL approaches depend on pre-trained preference models that output scalar rewards to approximate human preferences. However, these rewards often fail to capture human perception and are vulnerable to reward hacking, where higher scores do not correspond to better images. To address this, we introduce **Adv-GRPO**, an RL framework with an adversarial reward that iteratively updates both the reward model and the generator. The reward model is supervised using reference images as positive samples and can largely avoid being hacked. Unlike KL regularization that constrains parameter updates, our learned reward directly guides the generator through its visual outputs, leading to higher-quality images. Moreover, while optimizing existing reward functions can alleviate reward hacking, their inherent biases remain. For instance, PickScore may degrade image quality, whereas OCR-based rewards often reduce aesthetic fidelity. To address this, we take **the image itself as a reward**, using reference images and vision foundation models (e.g., DINO) to provide rich visual rewards. These dense visual signals, instead of a single scalar, lead to consistent gains across image quality, aesthetics, and task-specific metrics. Finally, we show that combining reference samples with foundation-model rewards enables distribution transfer and flexible style customization. In human evaluation, our method outperforms Flow-GRPO and SD3, achieving 70.0% and 72.4% win rates in image quality and aesthetics, respectively. Code and models have been released.

Date: November 25, 2025
Correspondence: Hao Chen at haochen.umd@gmail.com, Mike Zheng Shou at mike.zheng.shou@gmail.com,
Project Page: https://showlab.github.io/Adv-GRPO/

1 Introduction

Recently, online reinforcement learning (RL) has attracted increasing attention in large language models (LLMs) [6, 35, 50] and multimodal large language models (MLLMs) [4, 18, 26, 28, 37, 51]. In particular, the Group Relative Policy Optimization (GRPO) [35] algorithm, introduced by DeepSeek-R1 [6], has proven effective for aligning model behavior with reward signals in these domains. Motivated by these advances, several studies [5, 22, 40, 44, 45] have applied online RL to text-to-image (T2I) generation with diffusion models. For example, DanceGRPO [45] and Flow-GRPO [22] demonstrate that reward-driven optimization can improve performance when suitable reward models are provided.

However, despite these encouraging results, applying GRPO to T2I generation still faces fundamental challenges. The main difficulty lies in the misalignment between reward models and true human aesthetic preferences. In

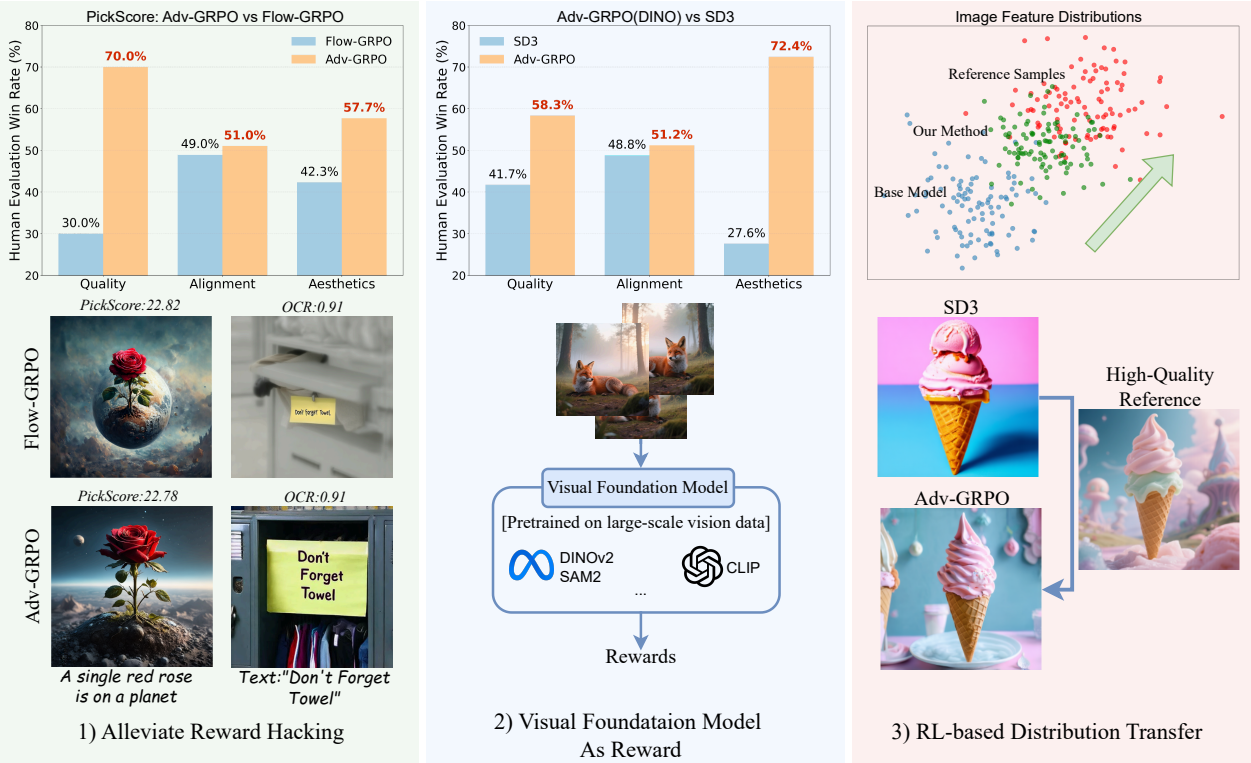

Figure 1 Overview of our approach. Our method Adv-GRPO improves text-to-image (T2I) generation in three ways: **1) Alleviate Reward Hacking**, achieving higher perceptual quality while maintaining comparable benchmark performance (e.g., PickScore, OCR), as shown in the top-left human evaluation panel; **2) Visual Foundation Model as Reward**, leveraging visual foundation models (e.g., DINO) for rich visual priors, leading to overall improvements as shown in middle-top human evaluation results; **3) RL-based Distribution Transfer**, enabling style customization by aligning generations with reference domains.

practice, many reward models produce scalar outputs that introduce biases toward specific visual attributes, such as oversaturated colors in CLIP-based, PickScore [15], or HPS [27, 43] models, or excessive text emphasis in OCR-based rewards. As a result, the generator may exploit these biases, achieving higher reward scores without genuine quality improvement, a phenomenon known as **reward hacking**. As shown in Fig. 2, Flow-GRPO underperforms the base model in several aspects, such as lower image quality with the PickScore reward and reduced aesthetics and quality under the OCR reward in human evaluation.

A common remedy is to add Kullback-Leibler (KL) regularization to contrain the parameter updates, which reduces reward hacking but limits optimization and lowers performance. To address this, we introduce **Adv-GRPO**, a novel RL framework with an adversarial reward that iteratively updates both the reward model and the base model. We observe that many high-quality reference images receive low scores from existing reward models. Therefore, we incorporate reference images as high-quality supervision, training the reward model as a discriminator to distinguish them from generated samples. Meanwhile, the base model as a generator is optimized with the GRPO loss. For the reward, we focus on human-preference reward models (e.g., HPS [27, 43], PickScore [15], Aesthetic models [7]), the main paradigm in T2I generation using our adversarial optimization. For other reward models, such as rule-based rewards (e.g., OCR), we leverage reference images in a multi-reward optimization scheme to enhance robustness. Our method consistently improves performance across both types, showing strong adaptability and generalization.

Although our method achieves better visual results and effectively mitigates reward hacking in existing reward models, some bias remains. For example, the PickScore reward tends to sacrifice image quality, while the OCR reward may reduce aesthetic fidelity. To address this, we directly use reference images as rewards by introducing a **reward model derived from a visual foundation model** [14, 29, 31, 33] to further optimize the T2I model. Specifically, we leverage the DINO [29] to provide stronger visual signals. Within our adversarial

optimization framework, DINO is fine-tuned as a reward model to guide the generator toward better visual alignment using reference samples. The output feature of DINO serves as the reward to optimize the base model. As a result, these dense visual signals, rather than a single scalar from the existing reward models, enable the base model to produce images with improved aesthetics, text alignment, and overall visual fidelity.

Moreover, we further introduce a new RL-based application for style transfer, where different reference datasets of distinct styles are used to effectively guide the base model toward specific visual domains. Finally, we conduct experiments across multiple benchmarks, and the results show that our method consistently improves image quality, text alignment, and aesthetics while maintaining comparable benchmark reward scores. Under the PickScore and OCR rewards, our method achieves winning rates of 70.0% and 85.3% in image aesthetics compared with Flow-GRPO in human evaluation. Under the DINO reward, it further achieves a 72.4% winning rate in aesthetics compared with the base model in human evaluation.

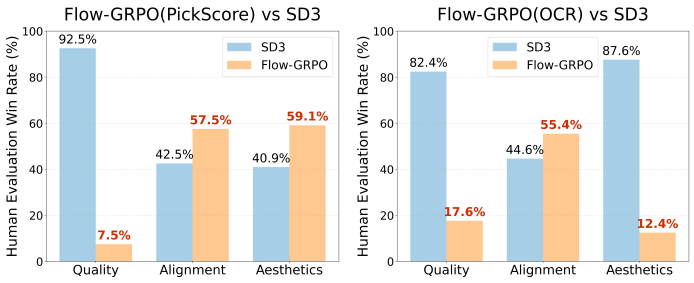

Figure 2 Human evaluation comparing Flow-GRPO and SD3 under PickScore and OCR rewards.

Our main contributions are summarized as follows:

- We are the first to introduce an RL framework with an adversarial reward that leverages high-quality reference images to jointly optimize the T2I model and reward model.
- We extend our approach to multiple types of existing reward models. Furthermore, we explore using visual foundation models as reward to guide the optimization of the base T2I model.
- Extensive experiments show that our method effectively alleviates reward hacking in existing reward models while maintaining competitive performance on standard benchmarks. With visual foundation model rewards, our method achieves comprehensive improvements in image quality, text–image alignment, and aesthetics.

2 Related Work

2.1 Reinforcement Learning for Image Generation

Recently, online reinforcement learning (RL) has shown strong effectiveness in improving the capabilities of large language models (LLMs) [6, 35, 50] and multimodal LLMs (MLLMs) [4, 18, 26, 28, 37, 51], and it has also been applied to text-to-image (T2I) generation. Compared with earlier RL methods such as PPO, GRPO is more efficient since it removes the need for an additional value network. In the context of T2I generation, prior methods such as DPO [8, 20, 23, 38, 46, 47] and PPO [3, 10, 12, 32, 49] have demonstrated effectiveness, and GRPO has also been adapted in this domain. For instance, DanceGRPO [45] applies GRPO to both image and video generation models [9, 16, 17, 30, 39], while Flow-GRPO [22] modifies the optimization process by replacing the ODE [24] with an SDE to improve sampling diversity. Despite these advances, such methods still face challenges including reward hacking and training instability. Building on Flow-GRPO, several work [19, 40] further refines the SDE process to stabilize optimization. Prior studies have sought to improve reward reliability in RL-based image generation. Works such as [1, 27, 48] reduce aesthetic bias through refined reward design, while SRPO [36] enhances efficiency using semantic positive–negative prompts. In contrast, we propose an adversarial training framework with reference samples.

2.2 Reward Models for Image Quality Assessment

In T2I generation, the main reward models are human-preference models, such as HPS [27, 43], HPDv2 [2], PickScore [15], and Aesthetic models [7, 34], which are built upon CLIP [31] and fine-tuned on large-scale human preference datasets. Other variants, like ImageReward [44] and UnifiedReward [41], further refine aesthetic alignment. In addition, rule-based rewards, such as OCR-based text accuracy and GenEval [11] for

object correctness, provide explicit but narrow supervision. However, both reward types are prone to reward hacking, as they often overfit to specific biases rather than capturing true perceptual quality.

3 Method

In this section, we first introduce the preliminaries of GRPO for flow matching and adversarial training in Sec. 3.1. We then describe our proposed adversarial optimization framework in Sec. 3.1. Finally, in Sec. 3.3, we extend our approach by leveraging a visual foundation model as the reward to further enhance overall image quality. An overview of the entire pipeline is illustrated in Fig. 3.

3.1 Preliminary

GRPO on Flow Matching. The denoising process in diffusion models [13, 21, 25] can be viewed as a Markov Decision Process (MDP), where each reverse step from x_t to x_{t-1} is treated as an action sampled from a policy $\pi_\theta(\cdot|x_t, c)$ conditioned on the current noisy sample x_t and the text prompt c. In practice, π_θ corresponds to the conditional flow distribution $p_\theta(x_{t-1}|x_t, c)$. At each iteration, we generate a group of G samples $\{x_0^i\}_{i=1}^G$ from the previous policy $\pi_{\theta_{\text{old}}}$ and compute their rewards $R(x_0^i, c)$. The group advantage $\hat{A}^i$ is obtained by normalizing the reward of each sample within the group:

$$\hat{A}^i = \frac{R(x_0^i, c) - \text{mean}\big(\{R(x_0^j, c)\}_{j=1}^G\big)}{\text{std}\big(\{R(x_0^j, c)\}_{j=1}^G\big)}. \tag{1}$$

GRPO then optimizes the policy model by maximizing the following objective:

$$f(r, \hat{A}, \theta, \epsilon, \beta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\min\big(r_t^i(\theta)\hat{A}^i,\ \tilde{r}_t^i(\theta)\hat{A}^i\big) - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot|x_t^i, c)\,\big\|\,\pi_{\theta_{\text{old}}}(\cdot|x_t^i, c)\big)\,, \tag{2}$$

with the importance ratio

$$r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i \mid x_t^i, c)}{p_{\theta_{\text{old}}}(x_{t-1}^i \mid x_t^i, c)}, \quad \tilde{r}_t^i(\theta) = \text{clip}\big(r_t^i(\theta), 1-\epsilon, 1+\epsilon\big). \tag{3}$$

Here, ϵ controls the clipping range and β weights the KL penalty to stabilize training.

Adversarial Training. Adversarial training is typically formulated as a minimax optimization problem between a generator G_θ and a discriminator D_ϕ:

$$\min_\theta \max_\phi\ \mathbb{E}_{x\sim p_{\text{data}}}\big[\log D_\phi(x)\big] + \mathbb{E}_{z\sim p_z}\big[\log\big(1 - D_\phi(G_\theta(z))\big)\big], \tag{4}$$

where p_{data} denotes the real data distribution and p_z is a prior distribution over latent variables. The discriminator D_ϕ is trained to distinguish real samples from generated ones, while the generator G_θ is optimized to produce samples that can fool the discriminator.

3.2 GRPO with Adversarial Reward

As shown in Fig. 3, our method extends GRPO to an adversarial setting, where the text-to-image (T2I) generator and the reward model are jointly optimized. The generator G_θ is trained via GRPO to maximize the rewards of its generated samples. The reward model R_ϕ serves as a discriminator, adversarially trained to distinguish high-quality reference images from generated ones. Specifically, given a text prompt c, the generator produces a group of samples $\{x_g^i = G_\theta(c)\}_{i=1}^G$ with corresponding reward values $R_\phi(x_g^i, c)$. The generator is optimized under the standard GRPO objective:

$$J_{\text{gen}}(\theta) = \mathbb{E}_{c\sim\mathcal{C},\{x_g^i\}_{i=1}^G\sim G_{\theta_{\text{old}}}}\Big[f(r, \hat{A}, \theta, \epsilon, \beta)\Big], \tag{5}$$

where $f(r, \hat{A}, \theta, \epsilon, \beta)$ follows the clipped GRPO formulation described in Eq. 2, and $\hat{A}$ denotes the normalized group advantage.

Meanwhile, the reward model is optimized using reference high-quality data $\mathcal{D}_{\text{ref}} = \{x_r\}$ as positive samples and generated images $\{x_g\}$ as negative samples:

$$J_{\text{reward}}(\phi) = -\mathbb{E}_{x_r \sim \mathcal{D}_{\text{ref}}}[\log R_\phi(x_r)] - \mathbb{E}_{x_g \sim G_\theta(c)}[\log(1 - R_\phi(x_g))]\,. \tag{6}$$

This adversarial co-training enforces the reward model to align with reference image distributions while guiding the generator toward higher-quality outputs.

This joint training objective establishes a dynamic equilibrium: the generator strives to produce images that maximize the reward, while the reward model continuously adapts by contrasting generated samples with high-quality references. In this process, reward hacking is effectively mitigated, as R_ϕ learns to better reflect perceptual quality beyond its initial biases, and G_θ is encouraged to improve both reward scores and overall visual fidelity.

Human Preference Models. Human preference models are the primary type of reward function used in current T2I generation. They are trained on human-labeled data, where annotators provide pairwise comparisons or aesthetic judgments to capture subjective visual preferences. We adopt such an adversarial co-training mechanism that leverages reference high-quality samples. Let $R(\cdot)$ denote the reward model, and let G denote the generator. We monitor the average reward scores for generated and reference samples respectively:

$$\bar{r}_{\text{gen}} = \mathbb{E}_{x_g \sim G}[R(x_g)], \quad \bar{r}_{\text{ref}} = \mathbb{E}_{x_r \sim D_{\text{ref}}}[R(x_r)]. \tag{7}$$

When the average reward of generated images surpasses that of reference images, i.e., $\bar{r}_{\text{gen}} > \bar{r}_{\text{ref}}$, we regard this as a signal of potential reward hacking. At this point, we trigger adversarial fine-tuning of the reward model, where reference samples are treated as positive samples and generated samples as negative samples. This process re-aligns the reward model toward human-preferred visual quality and prevents degenerate feedback loops during GRPO optimization. The optimization objective is defined in Eq. 6.

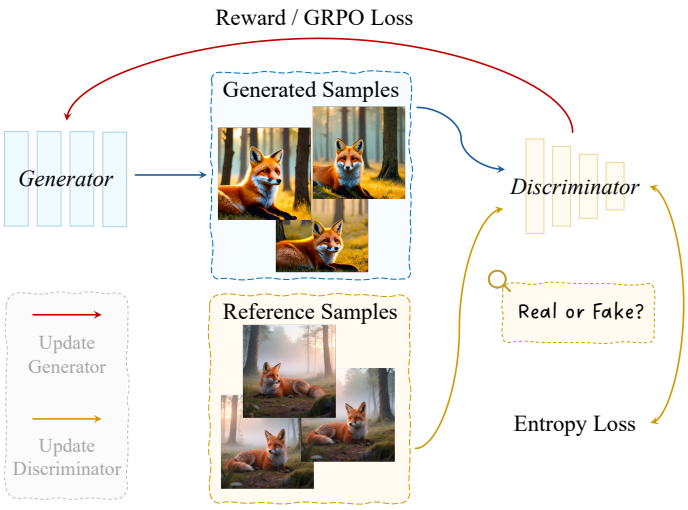

Figure 3 Pipeline of Adv-GRPO. The generator is optimized using the GRPO loss, while the discriminator is trained to distinguish between generated samples and reference images, treated as negative and positive samples, respectively. The discriminator serves as a reward model to provide feedback for the generator.

Rule-based Reward Models. Besides the main human-preference reward models, other rewards such as rule-based metrics (e.g., OCR or GenEval [11]) provide clear task-specific signals but are inherently deterministic and non-differentiable, making them unsuitable for adversarial training. To address this, we fully leverage high-quality reference images and adopt a simple multi-reward formulation to balance task specificity and visual realism. The reward is

$$R_{\text{combined}}(x_g, c) = \lambda\, R_{\text{rule}}(x_g, c) + (1 - \lambda)\, \text{sim}_{\text{CLIP}}(x_g, x_r), \tag{8}$$

where R_{rule} denotes the task-specific reward (e.g., OCR or GenEval), sim_{CLIP} measures the CLIP similarity between the generated image x_g and a reference image x_r, and $\lambda \in [0, 1]$ controls the trade-off. This formulation stabilizes training by preventing rule-based objectives from dominating and preserving overall visual fidelity.

3.3 Visual Foundation Models As Reward

The existing reward models provide explicit supervision but capture only limited aspects of image quality and often introduce aesthetic or content biases. Incorporating reference images via adversarial co-training alleviates reward hacking but mainly regularizes the reward model rather than holistically improving image quality.

Therefore, we further explore using **visual foundation models** as reward models to guide the optimization of the base generator. Unlike conventional reward models, visual foundation models encode rich semantic and structural representations of natural images, making them well-suited for aligning the overall distribution of generated images with that of high-quality reference images.

Formally, given a pre-trained visual foundation model $F_\psi(\cdot)$ (e.g., DINO [29]), we freeze its parameters and attach a lightweight binary classification head $h_\phi(\cdot)$ on top of its representations. For each input image x, we extract both the global [CLS] embedding and the patch-level features:

$$\mathbf{f}_{\text{cls}}, \mathbf{F}_{\text{patch}} = F_\psi(x), \tag{9}$$

where $\mathbf{f}_{\text{cls}} \in \mathbb{R}^D$ denotes the [CLS] token feature, and $\mathbf{F}_{\text{patch}} \in \mathbb{R}^{N \times D}$ represents the N patch embeddings.

Given the global [CLS] feature $\mathbf{f}_{\text{cls}}$ and patch-level features $\mathbf{F}_{\text{patch}} = \{\mathbf{f}_j\}_{j=1}^N$ extracted from the frozen visual backbone $F_\psi(\cdot)$, the reward is computed using a shared classification head $h_\phi(\cdot)$ as:

$$R_{\text{global}}(x) = h_\phi(\mathbf{f}_{\text{cls}}), \quad R_{\text{local}}(x) = \frac{1}{n}\sum_{j \in \mathcal{S}} h_\phi(\mathbf{f}_j), \tag{10}$$

where $\mathcal{S} \subset \{1, \ldots, N\}$ denotes a randomly selected subset of n patch tokens. This stochastic sampling encourages the model to focus on diverse local structures while maintaining computational efficiency. The final reward combines both components:

$$R_\phi(x) = \lambda_{\text{g}} R_{\text{global}}(x) + \lambda_{\text{l}} R_{\text{local}}(x), \tag{11}$$

where λ_{g} and λ_{l} control the relative contribution of global and local cues.

During adversarial training, the reward head h_ϕ is trained to discriminate reference images $x_r \sim \mathcal{D}_{\text{ref}}$ (positives) from generated images $x_g \sim G_\theta(c)$ (negatives). We employ a hinge loss objective for this discrimination. Specifically, h_ϕ is applied to both the global [CLS] feature and a subset of randomly sampled patch features extracted from the frozen backbone $F_\psi(\cdot)$, with separate hinge losses computed at the global and local levels. The corresponding hinge losses at the global and local levels are defined as:

$$\mathcal{L}_{\text{global}}(\phi) = \mathbb{E}_{x_r}[\max(0,\, 1 - h_\phi(\mathbf{f}^r_{\text{cls}}))] + \mathbb{E}_{x_g}[\max(0,\, 1 + h_\phi(\mathbf{f}^g_{\text{cls}}))]\,; \tag{12}$$

$$\mathcal{L}_{\text{local}}(\phi) = \mathbb{E}_{x_r}\left[\tfrac{1}{|\mathcal{S}|}\sum_{j \in \mathcal{S}} \max(0,\, 1 - h_\phi(\mathbf{f}^r_j))\right] + \mathbb{E}_{x_g}\left[\tfrac{1}{|\mathcal{S}|}\sum_{j \in \mathcal{S}} \max(0,\, 1 + h_\phi(\mathbf{f}^g_j))\right]. \tag{13}$$

The final adversarial loss for training the reward model is a weighted combination:

$$\mathcal{L}_{\text{reward}}(\phi) = \lambda_{\text{g}} \mathcal{L}_{\text{global}}(\phi) + \lambda_{\text{l}} \mathcal{L}_{\text{local}}(\phi), \tag{14}$$

where $\mathbf{f}^r_{\text{cls}}$ and $\mathbf{f}^g_{\text{cls}}$ denote the global features of reference and generated images, and $\mathbf{f}^r_j$, $\mathbf{f}^g_j$ represent their patch-level features.

This global–local reward formulation enables the generator to benefit from both complementary aspects: the global [CLS] feature emphasizes high-level semantics and structural consistency, while the local patch features capture fine-grained texture and detail. Together, they allow the model to generate more coherent and visually refined images.

4 Experiments

4.1 Implementation Details

Training Setup. We adopt Stable Diffusion 3 (SD3) [9] as the base generator. For the PickScore [15] reward, we use the PickScore prompt dataset for training and evaluation, and for the OCR reward, we employ the

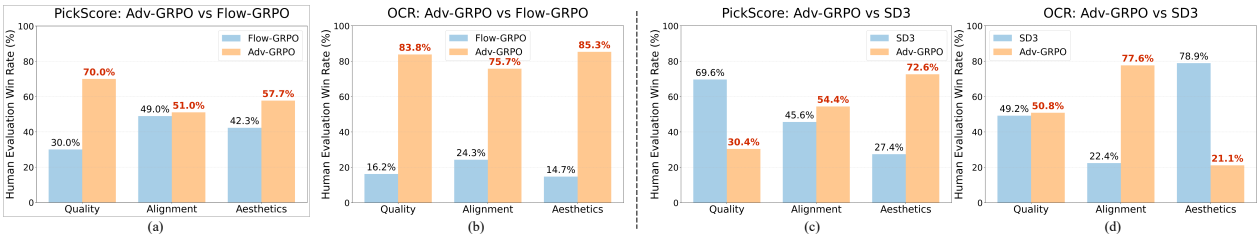

Figure 4 Human evaluation under PickScore- and OCR-based rewards. Our method Adv-GRPO improves image quality and aesthetics with PickScore reward in **a)**, and for all metrics with OCR reward in **b)**. Compared with the original model (SD3), PickScore reward trade-off aesthetic improvements with image quality degradation in **c)**, OCR reward trade-off text-alignment from aesthetics degradation in **d)**.

Figure 5 Visualizations under PickScore (**Left**) and OCR (**Right**) rewards. Our method Adv-GRPO alleviates reward hacking for both.

OCR prompt set. For visual foundation model experiments, we employ DINOv2 [29] as the reward model to optimize the base generator. Under the DINO [29] reward, our method is validated on PickScore, OCR, and GenEval prompts. Each prompt forms a group of 16 samples during training. We fine-tune only the last two layers of PickScore's vision branch (learning rate 3×10^{-4} for the generator and 5×10^{-6} for the reward model) for 1,000 iterations. For DINO, we train the classification head with a learning rate of 1×10^{-4}. In the OCR setting, we jointly optimize SD3 using both OCR and CLIP similarity rewards. Training uses 10 inference steps, with 2 timesteps randomly sampled from the 50–100% noise schedule. Eight reference images per prompt are generated with Qwen-Image [42]. All experiments are conducted on 8 NVIDIA H100 GPUs. Further details are provided in the supplementary material.

Baselines. We compare our method with two baselines: *Base Model*, the original SD3 without reinforcement learning optimization; and *Flow-GRPO* [22], a GRPO-based variant of SD3 that reformulates diffusion sampling as a stochastic differential equation to enhance training stability and diversity.

4.2 Evaluation Protocol

Metrics. We evaluate our method using these reward metrics: PickScore, OCR accuracy, and GenEval score [11]. In addition, we also compute the DINO similarity, which measures the cosine similarity between image embeddings extracted by the DINO, reflecting the semantic consistency between generated and reference images.

Human Evaluation. In addition, we conduct a comprehensive human evaluation covering three aspects: *Aesthetics*, *Alignment*, and *Quality*. The aesthetic score measures overall visual appeal and artistic composition. The alignment score evaluates the semantic consistency between generated images and text prompts, while the quality score reflects perceptual fidelity, structural coherence, and the absence of artifacts or distortions. We employ 12 expert evaluators to perform pairwise comparisons across 100 prompts for each reward setting, covering a total of 400 diverse prompts. In total, 10 comparison pairs across different rewards and methods are

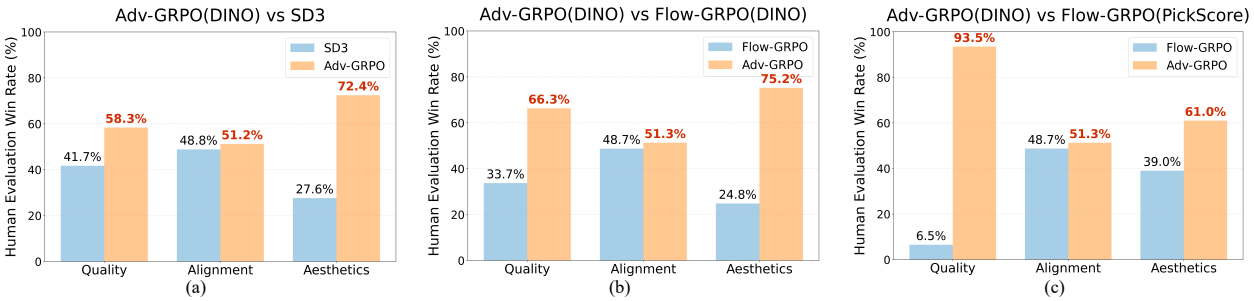

Figure 6 Human evaluation results under the visual foundation model (DINO) reward. Using a foundation model as the reward, our RL method improves image aesthetics, quality, and text alignment compared with the original SD3 model (a), and significantly outperforms Flow-GRPO under the DINO similarity reward (b) and PickScore reward (c).

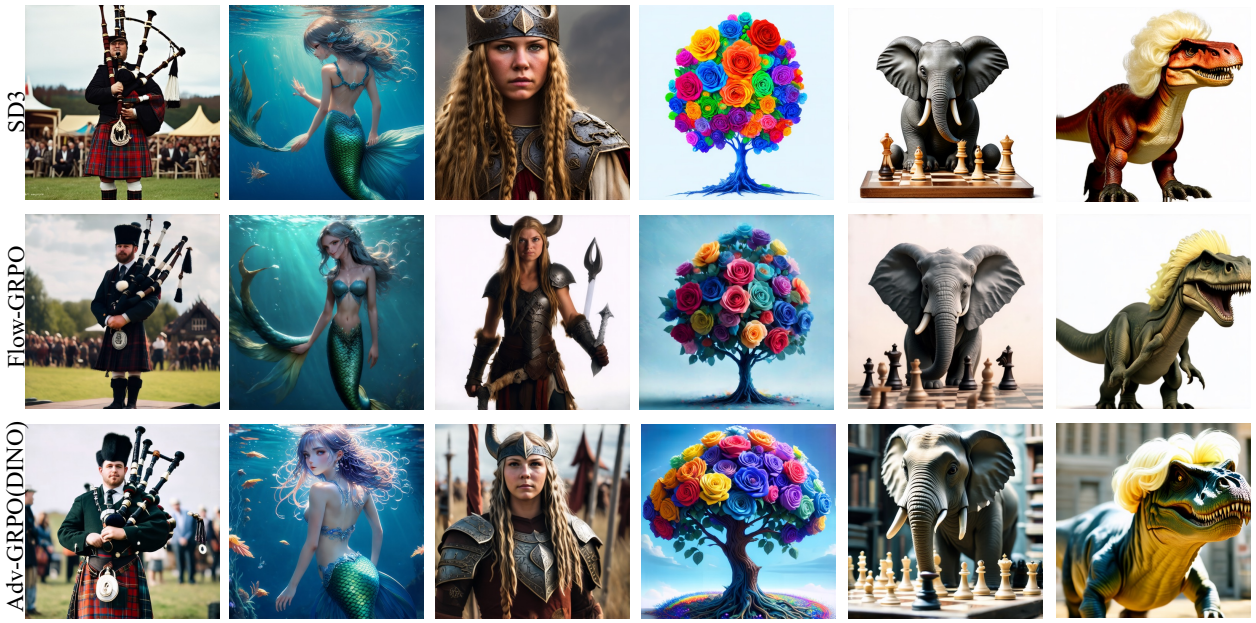

Figure 7 Visualizations under the DINO reward model. With **adversarial DINO reward**, our method shows better visual quality.

evaluated, resulting in 12,000 human comparison judgments. To ensure evaluation reliability, we conduct expert calibration, resolve inconsistent annotations, and continuously verify scoring criteria during the evaluation process. Further details of the evaluation protocol are provided in the supplementary material.

4.3 Main Results

Reward Hacking Mitigation.

We evaluate reward hacking from two perspectives. *a) Comparable benchmark performance with Flow-GRPO.* As shown in Tab. 1, our method achieves comparable benchmark scores to Flow-GRPO, indicating that adversarial training does not compromise quantitative performance in corresponding metrics. For PickScore, both methods reach around 22.80, substantially higher than 21.70 from SD3. For OCR, our method and Flow-GRPO achieve 0.91 accuracy, outperforming SD3 (0.58) by a large margin. This demonstrates that our approach maintains strong quantitative results while addressing reward bias.

Table 1 Comparison under different reward models. Each row corresponds to an independent optimization using the specific reward and its associated evaluation metric.

Reward Model	Metric	Method		
		SD3	Flow-GRPO	Adv-GRPO
PickScore	PickScore ↑	21.70	22.82	22.78
OCR	OCR Accuracy ↑	0.58	0.91	0.91

b) Significant improvement on human evaluation and visualizations. As shown in Fig. 4(a)(b), our method consistently outperforms Flow-GRPO under both PickScore and OCR rewards, achieving higher aesthetic, alignment, and quality scores. In particular, the win rate reaches 70% in image quality under PickScore and 85.3% in aesthetics under OCR. Compared with SD3 (Fig. 4(c)(d)), our method achieves a 72.6% win rate in aesthetics under the PickScore reward and a 77.6% win rate in alignment under the OCR reward, demonstrating substantial perceptual improvements. However, we also observe that PickScore optimization tends to sacrifice image quality, while OCR optimization slightly compromises aesthetics, indicating that some inherent bias in these reward models remains. Visualizations in Fig. 5 further confirm that our approach produces images with better perceptual quality and overall fidelity.

Vision Foundation Model as Better Rewards. We evaluate using DINO as reward models, and compare our method with Flow-GRPO (DINO similarity as the reward) and SD3. *a) Comprehensive improvements without degration.* Compared with SD3 in Fig. 6(a), our method consistently improves all human evaluation metrics, including aesthetics, alignment, and quality, especially the aesthetic dimension with 72.4% win rate. Compared with Flow-GRPO (using DINO similarity as the reward) in Fig. 6(b), our method achieves a 66.3% win rate in quality and a 75.2% win rate in aesthetics. Fig. 6(c) compares our method with Flow-GRPO (using PickScore reward), and our method achieves 93.5% win rate in quality. These results suggest that, compared with preference-based reward models, using a visual foundation model (DINO etc.) as the reward provides a more comprehensive and reliable guidance for image generation. The visualization results in Fig. 7 also show that our approach produces higher-quality images with richer backgrounds and improved aesthetics.

Figure 8 Visual comparison between our method (DINO reward) and SD3 across different task prompts.

b) Consistent improvement across benchmarks. We validate the versatility of the DINO reward using different benchmark prompts, including OCR and GenEval. As shown in Tab. 2, our adversarial DINO reward consistently improves performance across tasks, increasing OCR accuracy from 0.59 to 0.69 and GenEval score from 0.61 to 0.69 compared with SD3. The visual results in Fig. 8 also demonstrate visually appealing outputs, confirming that DINO serves as a general and reliable reward model across diverse objectives. Our reward curves exhibit a steady increase over training iterations, converging within roughly 1,000 steps, which is provided in the supplimentary material.

Table 2 General evaluation using the DINO reward across multiple tasks, comparing our method with SD3.

Method	PickScore ↑	OCR Accuracy ↑	GenEval ↑
SD3	21.70	0.59	0.61
Adv-GRPO (DINO)	21.90	0.69	0.69

4.4 Ablation Study

Number of Reference Images. We study the effect of the number of reference images by varying the dataset size across 200, 500, and 1,000 samples. As shown in Tab. 3, our method achieves comparable DINO similarity even with only 200 reference images, indicating that a small dataset is sufficient for effective optimization. The qualitative results in Fig. 10 further show that visual quality and style consistency remain stable as the number of references increases, demonstrating the data efficiency of our approach.

Comparison with Supervised Fine-Tuning (SFT). Human evaluation shows that our method under DINO reward model achieves notably higher perceptual quality than SFT, with over 70% win rates in both aesthetics and image quality in Fig. 6(d). As shown in Fig. 9 and Tab. 4, our approach also attains better visual results and higher quantitative metrics. Unlike SFT, which cannot explicitly optimize for specific reward objectives, our RL framework enables targeted optimization toward desired aspects such as text readability or visual appeal.

Figure 9 Ablation results. (a) Visualizations with different numbers of reference images, showing effectiveness even with 200 samples. (b) Visualizations of ablation studies on SFT, KL regularization, multi-reward optimization, and our method Adv-GRPO.

Table 3 Ablation on the number of reference samples used during inference. Our method maintains stable DINO similarity even with few reference images, demonstrating strong data efficiency.

Metric	SD3	w/ Fewer Reference Samples			Adv-GRPO
		200	500	1000	
DINO Similarity ↑	0.592	0.621	0.618	0.619	0.621

Table 4 Ablation on SFT, KL regularization, and multi-reward optimization under PickScore and OCR metrics.

Metric	SFT	Flow-GRPO (w/ KL)	Multi-Reward	Adv-GRPO
PickScore ↑	21.60	21.84	21.60	**22.78**
OCR Accuracy ↑	0.68	0.80	0.91	**0.91**

KL Regularization. We compare our method with Flow-GRPO using a KL regularization term. As shown in Tab. 4 and Fig. 9, adding a KL constraint leads to lower reward scores and degraded visual quality. KL regularization is sensitive, an overly large coefficient restricts optimization, while a small one cannot prevent reward hacking. Overall, our method achieves better stability and visual fidelity without relying on such fragile regularization.

Multi-Reward Combination. We also compare our method with Flow-GRPO trained using a combination of PickScore and OCR rewards. Although multi-reward optimization can also reduce reward hacking, balancing different reward weights is challenging due to varying sensitivities across models. As shown in Tab. 4 and Fig. 9, our method achieves higher metrics and better visual fidelity than the multi-reward baseline.

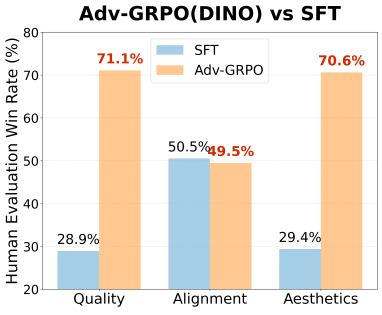

Figure 10 Human evaluation comparing our DINO-reward model with SFT, where our method performs better.

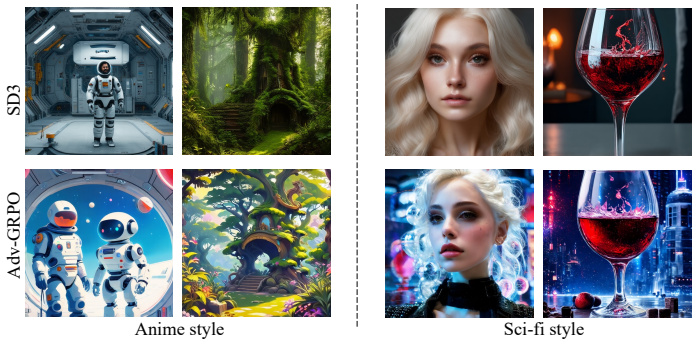

Figure 11 Application: Style transfer with the adversarial DINO reward. Our method successfully transfers the SD3 model to target visual domains, including **Anime** and **Sci-Fi** styles.

4.5 Style Customization via Adv-GRPO

We further demonstrate the versatility of our method through a style customization task. Unlike conventional RL-based T2I approaches that rely solely on text prompts and self-generated samples, our framework enables RL-driven optimization directly from pure image inputs, using visual foundation models such as DINO to guide learning. We construct two reference datasets, one anime-style and one sci-fi themed, and fine-tune the SD3 model with our proposed pipeline. As shown in Fig. 11, our method effectively transfers the generation style toward the reference domains while preserving semantic structure and image quality. This experiment showcases the flexibility and generalization of our approach, representing the first RL-based framework capable of performing style customization.

5 Conclusion

We introduce an RL framework with an adversarial reward for T2I generation. By leveraging reference high-quality references, the reward model better aligns with human visual preferences and mitigates reward hacking. Besides, incorporating visual foundation models such as DINO further provides unbiased visual guidance, improving overall image quality, aesthetics and text alignment. Extensive experiments verify the effectiveness and generality of our framework across diverse reward settings.

6 Acknowledgement

We thank Danze Chen and Kaiming Yang for their support in reference data generation. We are also grateful to Jiaming Han, Yuang Ai, and Shaobin Zhuang for their valuable advice and insightful discussions.

References

[1] Ying Ba, Tianyu Zhang, Yalong Bai, Wenyi Mo, Tao Liang, Bing Su, and Ji-Rong Wen. Enhancing reward models for high-quality image generation: Beyond text-image alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19022–19031, 2025.

[2] Ying Ba, Tianyu Zhang, Yalong Bai, Wenyi Mo, Tao Liang, Bing Su, and Ji-Rong Wen. Enhancing reward models for high-quality image generation: Beyond text-image alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19022–19031, 2025.

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

[4] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. arXiv preprint arXiv:2504.11468, 2025.

[5] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400, 2023.

[6] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR, abs/2501.12948, 2025.

[7] discus0434. Aesthetic predictor v2.5. https://github.com/discus0434/aesthetic-predictor-v2-5, 2025. Accessed: 2025-06-10.

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

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In ICML, 2024.

[10] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023. Neural Information Processing Systems Foundation, 2023.

[11] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In NeurIPS, 2023.

[12] Shashank Gupta, Chaitanya Ahuja, Tsung-Yu Lin, Sreya Dutta Roy, Harrie Oosterhuis, Maarten de Rijke, and Satya Narayan Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. arXiv preprint arXiv:2503.00897, 2025.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NeurIPS, pages 6840–6851, 2020.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In ICCV, pages 4015–4026, 2023.

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems, 36:36652–36663, 2023.

[16] W Kong, Q Tian, Z Zhang, R Min, Z Dai, J Zhou, J Xiong, X Li, B Wu, J Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv. org/abs/2412.03603.

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[18] Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. arXiv preprint arXiv:2509.07969, 2025.

[19] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802, 2025.

[20] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. arXiv preprint arXiv:2406.04314, 2(5):7, 2024.

[21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

[22] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470, 2025.

[23] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. arXiv preprint arXiv:2501.13918, 2025.

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022.

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022.

[26] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv. org/abs/2503.20783.

[27] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025a. URL https://arxiv. org/abs/2508.03789.

[28] Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning. arXiv preprint arXiv:2505.23380, 2025.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

[30] 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. arXiv preprint arXiv:2307.01952, 2023.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, pages 8748–8763, 2021.

[32] 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 NeurIPS, 2023.

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.

[34] Christoph Schuhmann. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2023-11-10.

[35] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. CoRR, abs/2402.03300, 2024.

[36] Xiangwei Shen, Zhimin Li, Zhantao Yang, Shiyi Zhang, Yingfang Zhang, Donghao Li, Chunyu Wang, Qinglin Lu, and Yansong Tang. Directly aligning the full diffusion trajectory with fine-grained human preference. arXiv preprint arXiv:2509.06942, 2025.

[37] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966, 2025.

[38] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8228–8238, 2024.

[39] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[40] Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. arXiv preprint arXiv:2509.05952, 2025.

[41] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236, 2025.

[42] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025.

[43] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2096–2105, 2023.

[44] 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, 36:15903–15935, 2023.

[45] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818, 2025.

[46] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model (2024). URL https://arxiv.org/abs/2311.13231.

[47] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. Advances in Neural Information Processing Systems, 37:73366–73398, 2024.

[48] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8018–8027, 2024.

[49] Hanyang Zhao, Haoxian Chen, Ji Zhang, David D Yao, and Wenpin Tang. Score as action: Fine-tuning diffusion generative models by continuous-time reinforcement learning. arXiv preprint arXiv:2502.01819, 2025.

[50] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. arXiv preprint arXiv:2507.18071, 2025.

[51] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362, 2025.

Appendix

In this supplementary material, we provide additional visualization results in Sec. B, further style transfer examples in Sec. C, experiments using SigLIP for optimization in Sec. D, more implementation details in Sec. A, the reward curves in Sec. E, and the full procedures of our human evaluation in Sec. F.

A More Implementation Details

In the DINO reward, we assign a 7:3 weighting ratio to the global and local batch losses and rewards. For SD3, we apply LoRA-based fine-tuning with a configuration that uses a rank of 32, a scaling factor (lora_alpha) of 64, and Gaussian initialization for all LoRA weights. During both training and evaluation, we set the classifier-free guidance (CFG) scale to 4.5, and employ bfloat16 mixed precision throughout the process. For the DINO reward training schedule, we adopt a 10:1 update ratio, meaning that the discriminator is updated for 10 steps for every 1 generator step. For the PickScore reward model, we perform fine-tuning only when the reward assigned to the generated images surpasses that of the reference images.

B Visualizations Under Our Method

Alleviating Reward Hacking. We provide additional visualizations to further demonstrate the effectiveness of our method across various reward models. As shown in Fig. 12, our approach significantly alleviates reward hacking issues present in existing reward models such as PickScore and OCR, producing images with consistently higher overall visual quality compared with Flow-GRPO.

More Visualizations under DINO reward. In addition, Fig. 13 presents more visualization results obtained under the adversarial DINO reward model. These results show that our method generates images with stronger compositional quality, richer color saturation, improved aesthetic appeal, and more diverse background details, further validating the robustness and generalization ability of our approach.

C More Visualizations on Style Customization

As shown in Fig. 16, our method successfully transfers the base model's style to an anime style using anime reference images. These results demonstrate that our RL-based approach, guided by a visual foundation model, can effectively achieve style customization.

D Using SigLIP for Optimization

As shown in Fig. 15, in addition to DINO, we also experiment with SigLIP as the visual foundation model used for optimization. The pipeline follows the same structure as DINO: we attach a lightweight head to SigLIP and use it to classify reference images and generated images. In this setup, SigLIP serves as the discriminator, while SD3 functions as the generator. Unlike DINO, which provides both global and local features, SigLIP offers only global representations. The successful performance under SigLIP demonstrates that **our method generalizes well to visual foundation models beyond DINO.**

E Reward Curve

We report the reward curve obtained during training, as illustrated in Fig. 17. The results show that training converges within approximately 1000 steps. In addition, the reward of our generated images consistently surpasses that of the reference images (produced by the QWen model) throughout the training process.

F Human Evaluation

For the human evaluation, we assess model performance across three dimensions: *image quality*, *image aesthetics*, and *text-image alignment*. For each question, experts are presented with two images generated by two different models and are asked to select the better one along all three dimensions, as shown in Fig. 18.

Figure 12 More Visualizations about alleviating reward hacking under PickScore and OCR reward models.

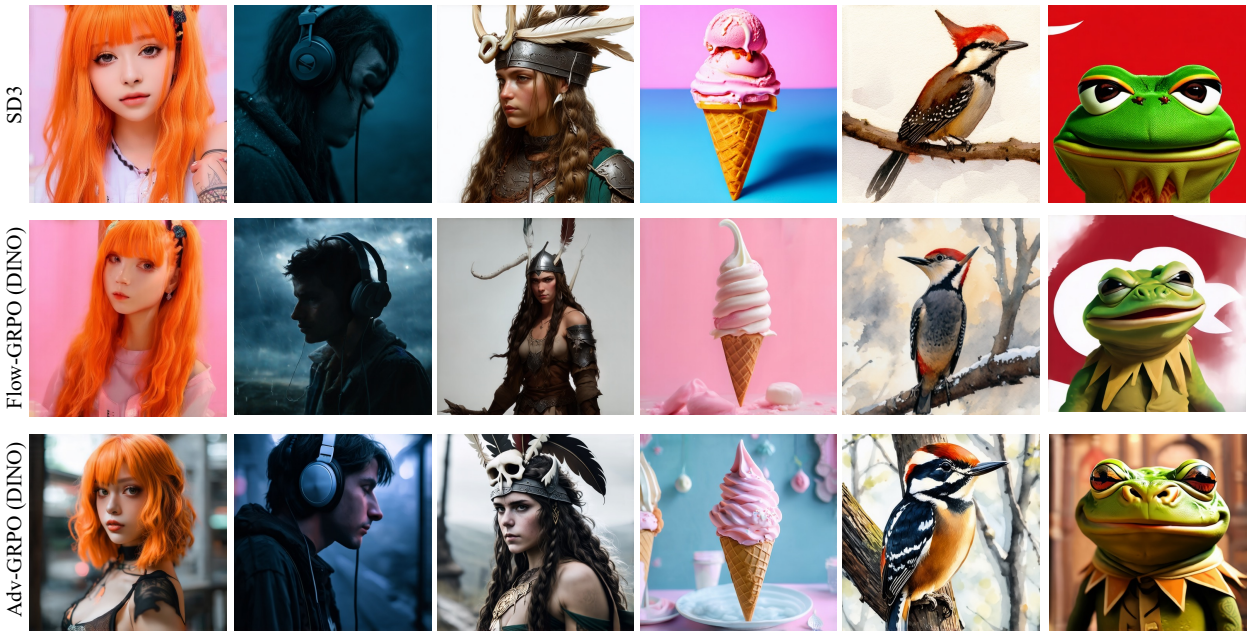

Figure 13 Additional visualizations using the DINO reward model. Our method produces images with consistently higher visual quality.

Figure 14 More visualizations with DINO reward using different benchmark OCR and GenEval prompts.

We construct a benchmark consisting of **10 groups** of comparison tasks, with a total of **100 questions**. Each group is evaluated by **12 experts**, and each question receives annotations from **3 independent experts**. This setup results in **300 individual annotation data points** (100 questions $\times$ 3 annotators per question), from which

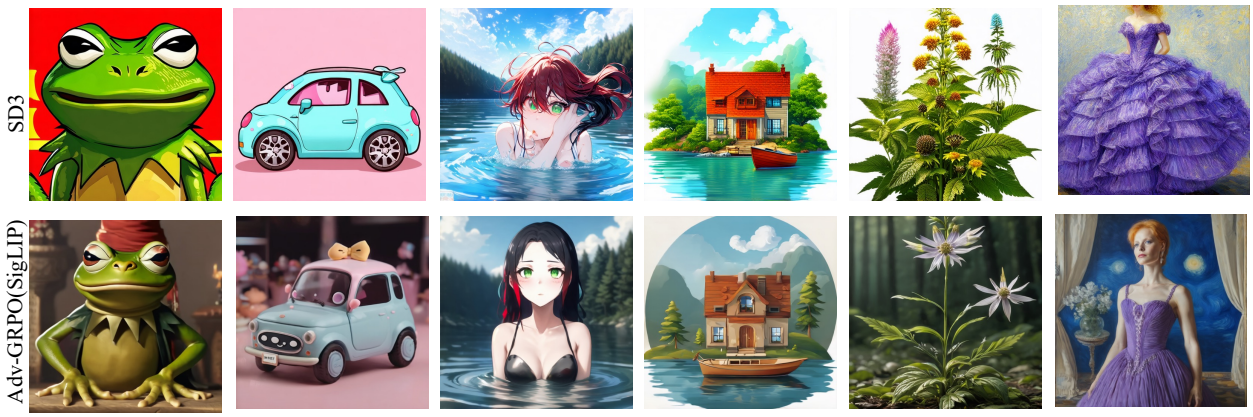

Figure 15 Visualizations with the SigLIP reward. Compared with SD3, using other visual foundation models such as SigLIP as the reward function can also lead to overall improvements in image quality.

Figure 16 More style customization results. Using anime reference images, our method effectively transfers the base model's style to an anime aesthetic, guided by the provided samples.

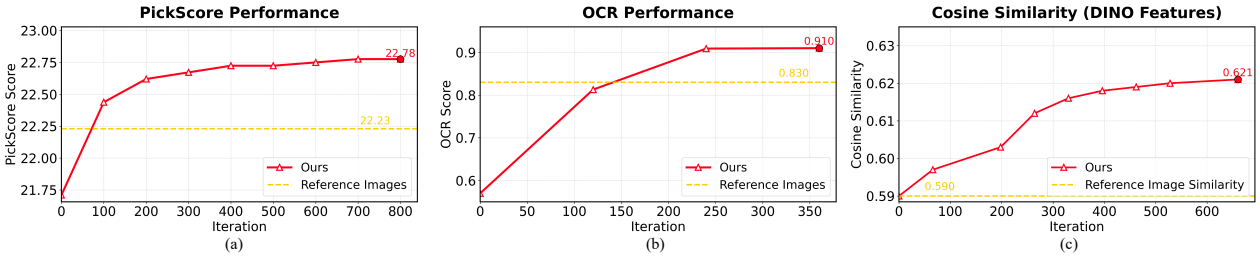

Figure 17 Reward curves under different reward models. We shows the training dynamics of our method and the baseline under three reward models: (a) PickScore, (b) OCR accuracy, and (c) DINO cosine similarity.

we derive the final aggregated results.

To ensure the reliability of the human evaluation, we adopt a multi-step quality-control protocol. First, we conduct **expert calibration**, during which annotators review reference examples and align on the scoring criteria. During the evaluation, we monitor and **resolve inconsistent annotations** through cross-checking and adjudication when needed. In addition, we **continuously verify and refine the scoring guidelines** throughout the evaluation to minimize ambiguity and ensure consistent interpretation across annotators.

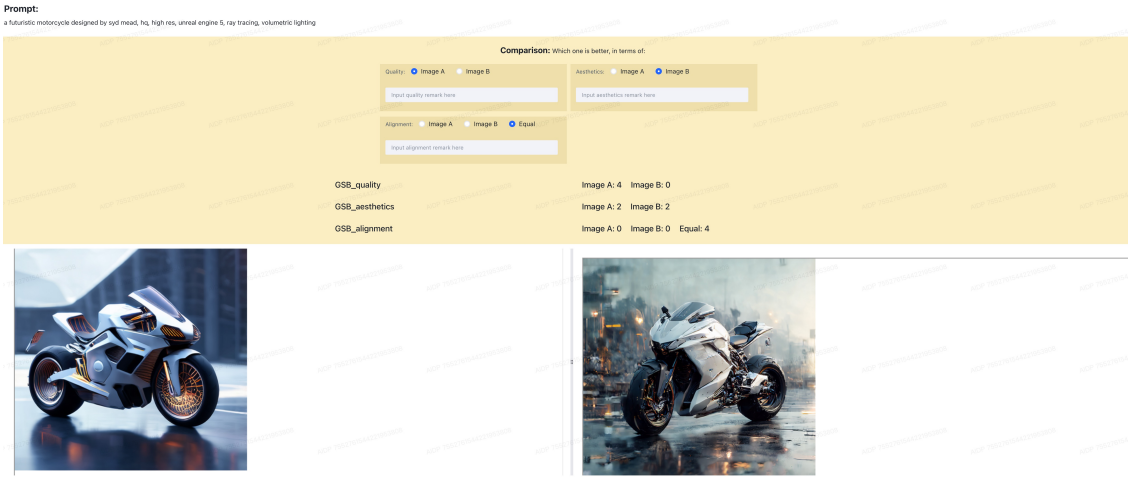

Figure 18 Screenshot of the interface used in our human evaluation study.