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

Supplementary Material

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.

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

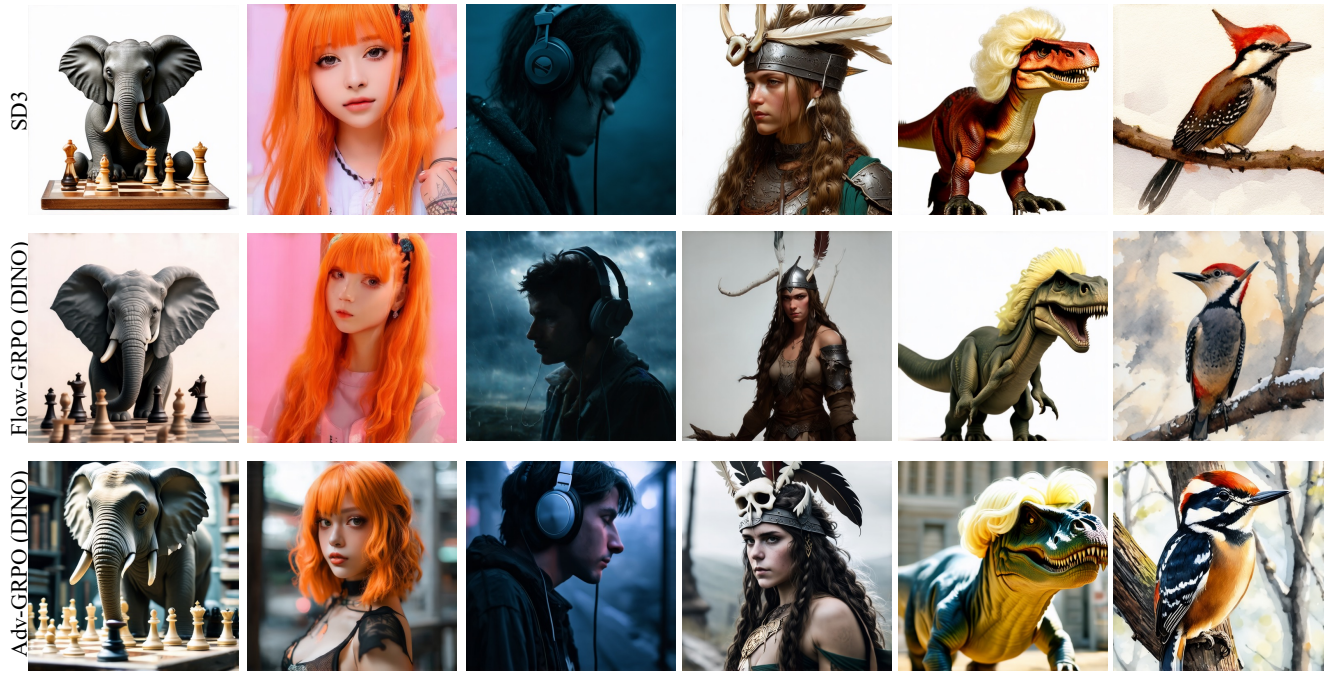

Figure 13. Additional visualizations using the DINO reward model. Our method produces images with consistently higher visual quality.

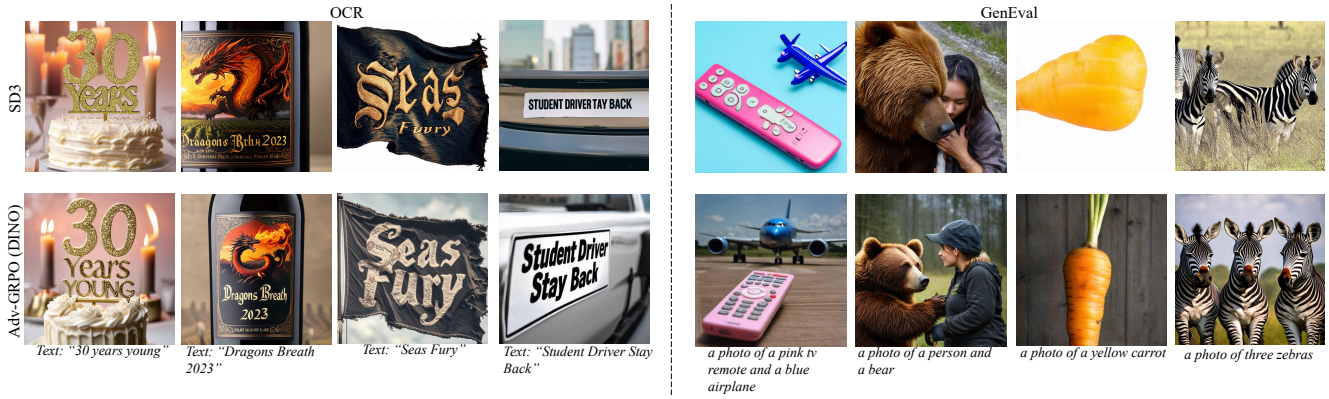

Figure 14. More visualizations with DINO reward using different benchmark OCR and GenEval prompts.

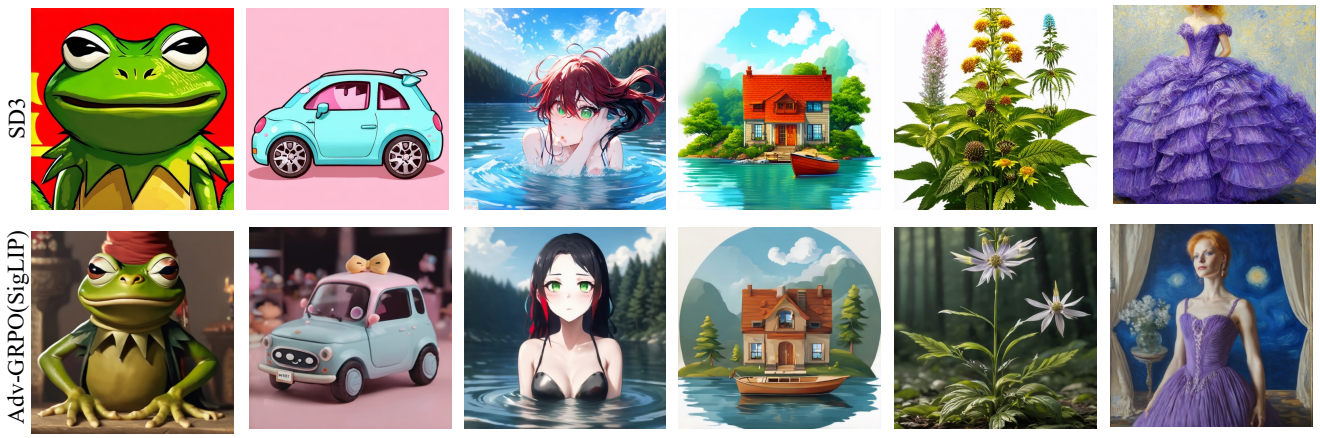

Figure 15. Visualizations with the SigLIP reward. Compared with SD3, using other visual foundation models such as SigLIP as the reward function can also lead to overall improvements in image quality.

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.

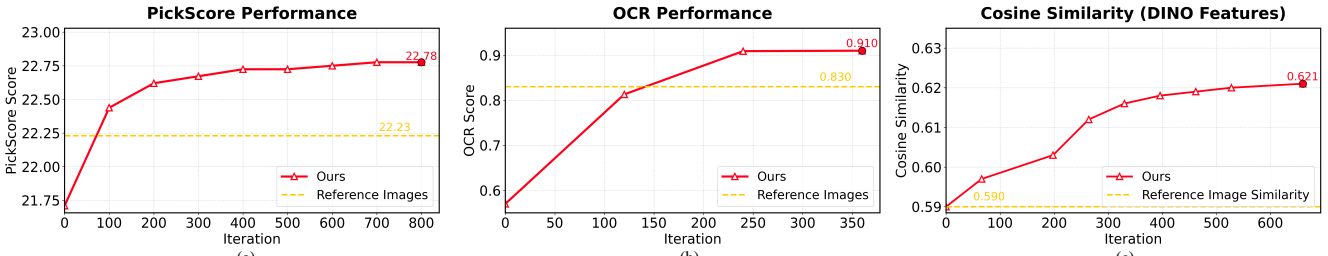

Figure 17. **Reward curves under different reward models.** We shows the training dynamics of our method and the baseline under three reward models: (a) PickScore, (b) OCR accuracy, and (c) DINO cosine similarity.

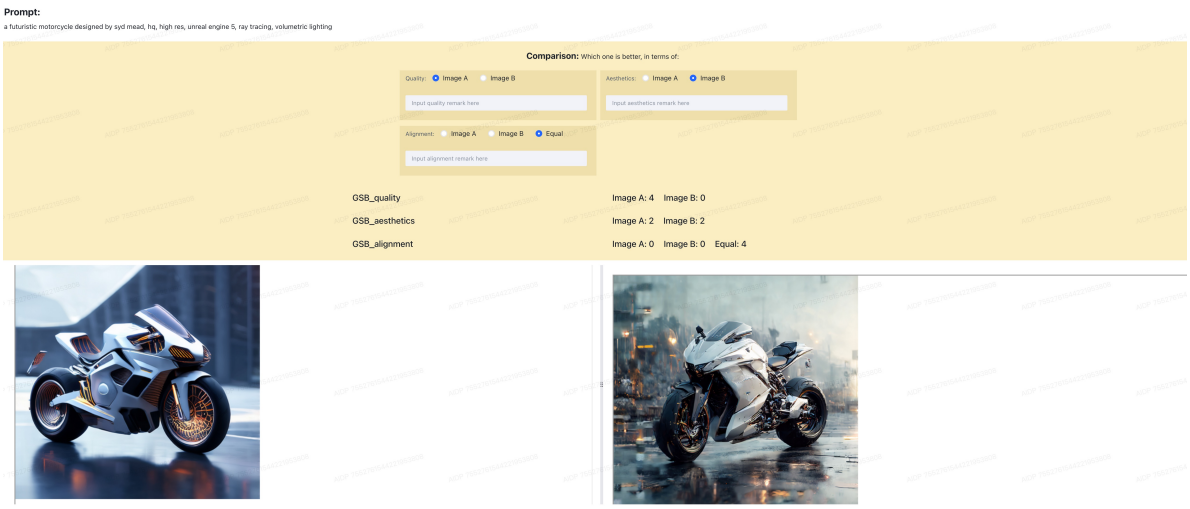

Figure 18. Screenshot of the interface used in our human evaluation study.