GDRO: Group-level Reward Post-training Suitable for Diffusion Models

Supplementary Material

In the appendix, we first provide more theoretical analysis on some proofs that we omitted in the main body for simplicity in Sec. A. Then we provide more experiment results in Sec. B, including more ablation studies in Sec. B.1 and more visualizations in Sec. B.2.

A. More Theoretical Analysis

A.1. Implicit Reward Function

In the main body, we introduce the implicit reward function proposed by DPO methods:

$$s_\theta(x) = r(x_0, c) = \beta_{\text{KL}} \log \frac{\pi^*_\theta(x_0 \mid c)}{\pi_{\text{ref}}(x_0 \mid c)} + \underbrace{\beta_{\text{KL}} \log Z(c)}_{\text{can be deleted}} \tag{13}$$

The scalar term $C = \beta \log Z(c)$ of this implicit function can be eliminated during the calculation toward the DPO objective because

$$\begin{aligned}
&\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma\big(s_\theta(x_0^+, c) - s_\theta(x_0^-, c)\big) \\
&= -\log \sigma\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_0^+ \mid c)}{\pi_{\text{ref}}(x_0^+ \mid c)} + C - \beta_{\text{KL}} \log \frac{\pi^*_\theta(x_0^- \mid c)}{\pi_{\text{ref}}(x_0^- \mid c)} - C\big) \\
&= -\log \sigma\big[\beta_{\text{KL}}\big(\log \frac{\pi^*_\theta(x_0^+ \mid c)}{\pi_{\text{ref}}(x_0^+ \mid c)} - \log \frac{\pi^*_\theta(x_0^- \mid c)}{\pi_{\text{ref}}(x_0^- \mid c)}\big)\big].
\end{aligned} \tag{14}$$

This scalar term can also be canceled in the derivations toward the GDRO objective. Now we prove this from the Plackett-Luce ranking model perspective. According to the PL model, with the images $x_1, ..., x_k$ and their corresponding explicit rewards $r_1, ..., r_k$, the likelihood of the ranking $x_1 \succ ... \succ x_k$ is:

$$P(x_1 \succ ... \succ x_k) = \prod_{i=1}^{k} \frac{\exp\big(r(x_i, c)\big)}{\sum_{j=i}^{k} \exp\big(r(x_j, c)\big)}. \tag{15}$$

The likelihood of this ranking computed from the implicit reward function is then

$$\begin{aligned}
P(x_1 \succ ... \succ x_k) &= \prod_{i=1}^{k} \frac{\exp\big(s(x_i, t)\big)}{\sum_{j=i}^{k} \exp\big(s(x_j, t)\big)} \\
&= \prod_{i=1}^{k} \frac{\exp\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_i|c)}{\pi_{\text{ref}}(x_i|c)} + C\big)}{\sum_{j=i}^{k} \exp\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_j|c)}{\pi_{\text{ref}}(x_j|c)} + C\big)} \\
&= \prod_{i=1}^{k} \frac{\exp(C) \exp\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_i|c)}{\pi_{\text{ref}}(x_i|c)}\big)}{\exp(C) \sum_{j=i}^{k} \exp\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_j|c)}{\pi_{\text{ref}}(x_j|c)}\big)} \\
&= \prod_{i=1}^{k} \frac{\exp\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_i|c)}{\pi_{\text{ref}}(x_i|c)}\big)}{\sum_{j=i}^{k} \exp\big(\beta_{\text{KL}} \log \frac{\pi^*_\theta(x_j|c)}{\pi_{\text{ref}}(x_j|c)}\big)}
\end{aligned} \tag{16}$$

The scalar term C is therefore canceled before the following derivations toward the GDRO objective.

A.2. GDRO and DPO

The objective of GDRO is:

$$\mathcal{L}_{\text{GDRO}}(\theta) = \sum_{i=1}^{k-1} \left(\log \sum_{m=i}^{k} \exp\big(s_\theta(x_m, t)\big) - \sum_{j=i}^{k} q_i(j, \tau) s_\theta(x_j, t) \right). \tag{17}$$

When $\tau \to 0$, $q_i(j, \tau) = 1$ if $i = j$, and $q_i(j, \tau) = 0$ if $i \neq j$. Therefore, when $k = 2, \tau \to 0$ the objective becomes:

$$\begin{aligned}
\lim_{\tau \to 0} \mathcal{L}^{k=2}_{\text{GDRO}}(\theta) &= \sum_{i=1}^{2-1} \left(\log \sum_{m=i}^{2} \exp\big(s_m\big) - \sum_{j=i}^{2} q_i(j, \tau) s_j \right) \\
&= \log \sum_{m=1}^{2} \exp\big(s_m\big) - \sum_{j=1}^{2} q_i(j, \tau) s_j \\
&= \log\big(e^{s_1} + e^{s_2}\big) - s_1 \\
&= \log\big(e^{s_1}(1 + e^{s_2 - s_1})\big) - s_1 \\
&= \log\big(1 + e^{-\Delta}\big),
\end{aligned} \tag{18}$$

where $s_i = s_\theta(x_i, t)$, $\Delta = s_1 - s_2$.

Given that $\sigma(x) = \frac{1}{1+e^{-x}}$, we can reformulate the above equation into:

$$\begin{aligned}
\lim_{\tau \to 0} \mathcal{L}^{k=2}_{\text{GDRO}}(\theta) &= \log\big(1 + e^{-\Delta}\big) \\
&= -\log\big(\frac{1}{1 + e^{-\Delta}}\big) \\
&= -\log \sigma(\Delta) \\
&= -\log \sigma\big(s_1 - s_2\big) = \mathcal{L}_{\text{DPO}}.
\end{aligned} \tag{19}$$

This means that, when $k = 2, \tau \to 0$, GDRO reduces to DPO. This is because when $\tau \to 0$, the explicit reward is omitted during computation, making only the ranking matter. And when $k = 2$, the ranking turns into only pair-wise preference. From this perspective, we can view GDRO as a natural extension of DPO-based methods that not only enlarge the group size from 2 to any k, but also consider the information of the explicit rewards instead of preferences.

B. More Experiments

B.1. Ablation Studies

Beta β. We have provided the ablation study on beta in the main body. Now we provide visualizations on the case where $\beta = 6$ in the OCR task in Fig. vii. In the main body, the evaluation curve of $\beta = 6$ in OCR is deceptive, as the original reward gets very high at an early optimization step. However, we show that a collapse happens for this case using the corrected score. Now we present the images showing what collapsed images look like, even if they earn a

Figure vii. **Demosntration on collapse**. When $\beta = 6$, though an evaluation time reward of 0.85 on OCR is achieved, the images actually collapse, which is another case of reward hacking.

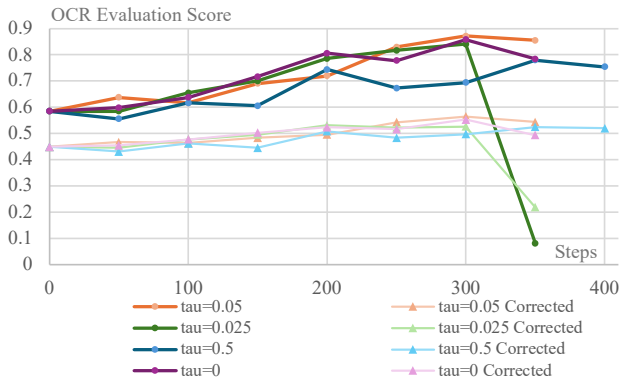


Figure viii. **Ablation study on temperature**. We provide the evaluation curves on different choices of the temperature τ.

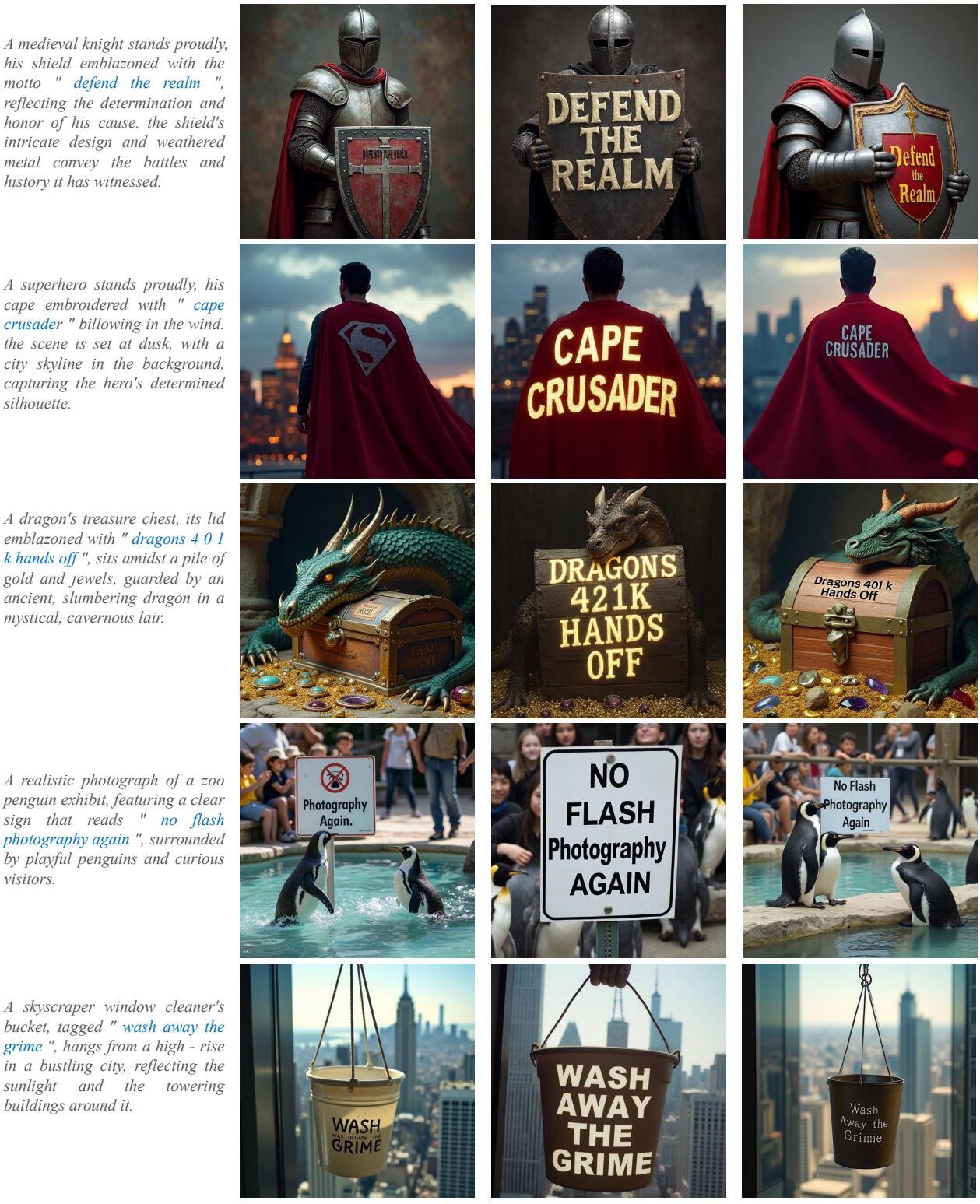
high evaluation time reward. In Fig. vii, these are evaluation images of $\beta = 6$ when it achieves an evaluation reward of 0.85 in the OCR task at step 100. Though the number looks promising, the images demonstrate noticeable stripe artifacts as well as severe quality degradation. In this case, though the texts look right, the generation quality of the model is greatly compromised, making this optimization a failure. This further warns us that only looking at the evaluation reward is improper and deceptive.

Note that without top-1 likelihood stabilization, the optimization process will have similar collapse issues like these images, demonstrating a high reward but degraded quality.

Temperature τ. The temperature τ is a hyperparameter used to compute the explicit reward distribution $q(i, \tau) = \text{softmax}(\frac{r_i}{\tau})$. This temperature controls how peaky this distribution is. When τ is bigger, the distribution is more even and the differences between different reward values are reduced. When τ is smaller, the differences of the reward values are amplified during calculation. We perform the ablation study on the choice of τ on the OCR task. With the chosen values of $0.5, 0.05, 0.025$ and 0. Note that for the $\tau = 0$ here, the explicit rewards are omitted and the explicit reward distribution reduces to a one-hot vector. For other hyperparameters, we use $\gamma = 0.5, lr = 3e-4, \beta = 12, k = 8$. The evaluation curves regarding τ are shown in Fig. viii. A higher temperature $\tau = 0.5$ leads to insufficient optimization, as both the original scores and the corrected score curves are all below others. We posit that this is because a higher temperature lowers the differences between different rewards, making the model not sufficiently learns from the advantages of high-reward samples. A lower temperature $\tau = 0.025$ leads to unstable optimization, as the curves collapse. A medium choice of temperature $\tau = 0.05$ leads to the most satisfying result, demonstrating a good original reward as well as a good corrected curve, showing the best stability. A special case is where $\tau = 0$, which stands for the case without explicit reward. In this case, GDRO gets a nice performance between $\tau = 0.05$ and $\tau = 0.025$, showing that the ranking itself can provide enough information to let the model learn the distinctions between good samples and bad samples. But still, providing more information using explicit rewards can bring more improvements.

B.2. More Visualizations

We provide more visualizations of GDRO and Flow-GRPO. In Fig. ix, Fig. x, we provide more cases on the OCR task. Similar to the analysis in the main body, Flow-GRPO tends to enlarge, bold, and rectify the text orientations to make the text more readable to the OCR model, and thus making other elements get omitted or even disappear. This reward hacking case is not the same as the collapse one in Fig. vii that directly degrades the quality because of collapse, but gradually makes the details of the image fewer and thereby lowering the overall quality. In contrast, GDRO doesn't have such a trend, making the images show correct texts while being abundant in details as described by the text prompts. In Fig. xi and Fig. xii, we provide more cases on the GenEval task. As shown in the image, Flow-GRPO images mostly reduce to flat-drawing-like images with a plain background and disproportionate objects, ignoring most of the details. In contrast, GDRO demonstrates good visual quality as well as correct object attributes, showing its robustness in mitigating reward hacking.



FLUX.1(0.58) Flow-GRPO (*0.87*) **Ours (*0.87*)**

Figure ix. **More visualizations on OCR**. We provide more comparisons between our method and Flow-GRPO when the evaluation reward is the same on the OCR task.

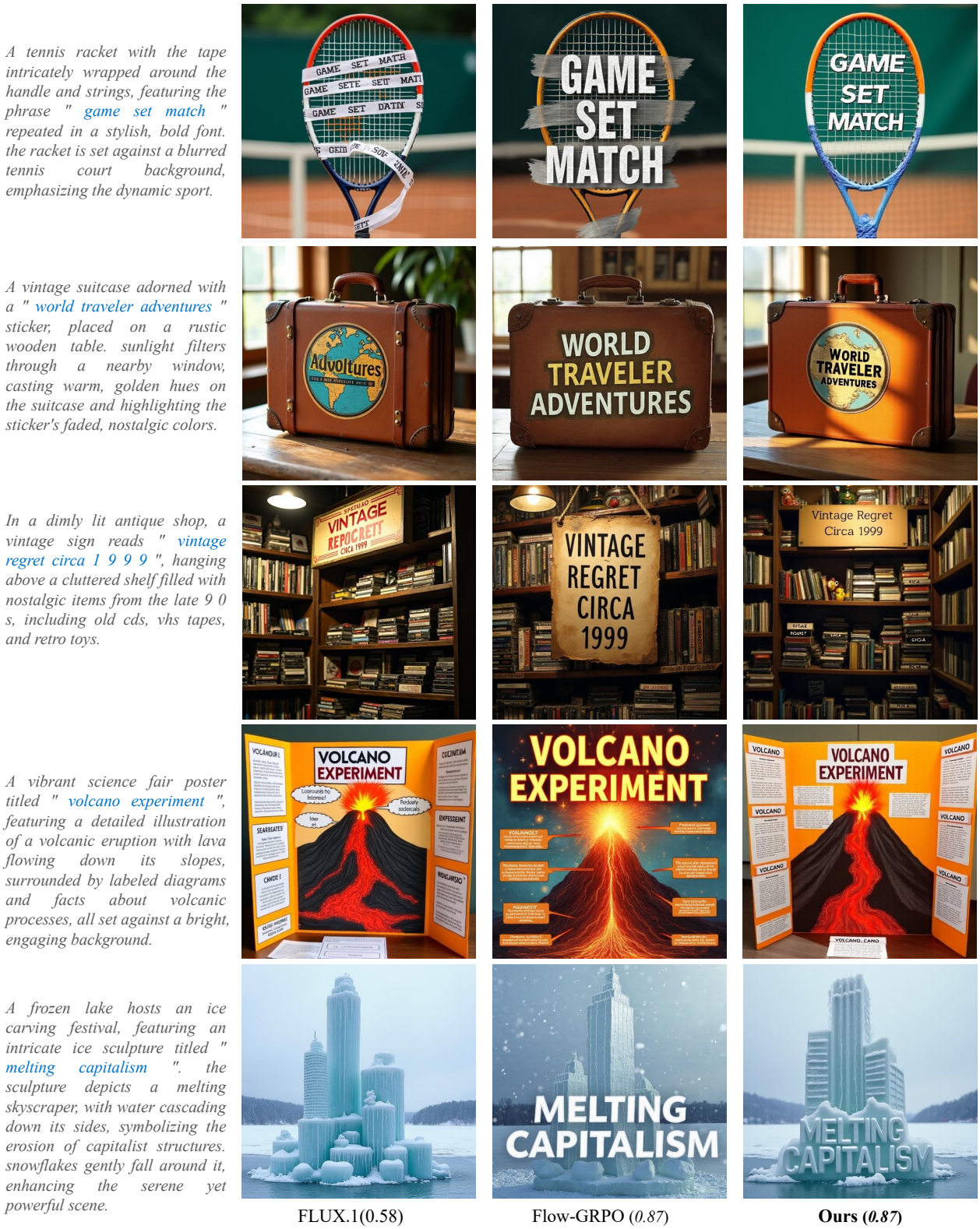


Figure x. **More visualizations on OCR**. We provide more comparisons between our method and Flow-GRPO when the evaluation reward is the same on the OCR task.

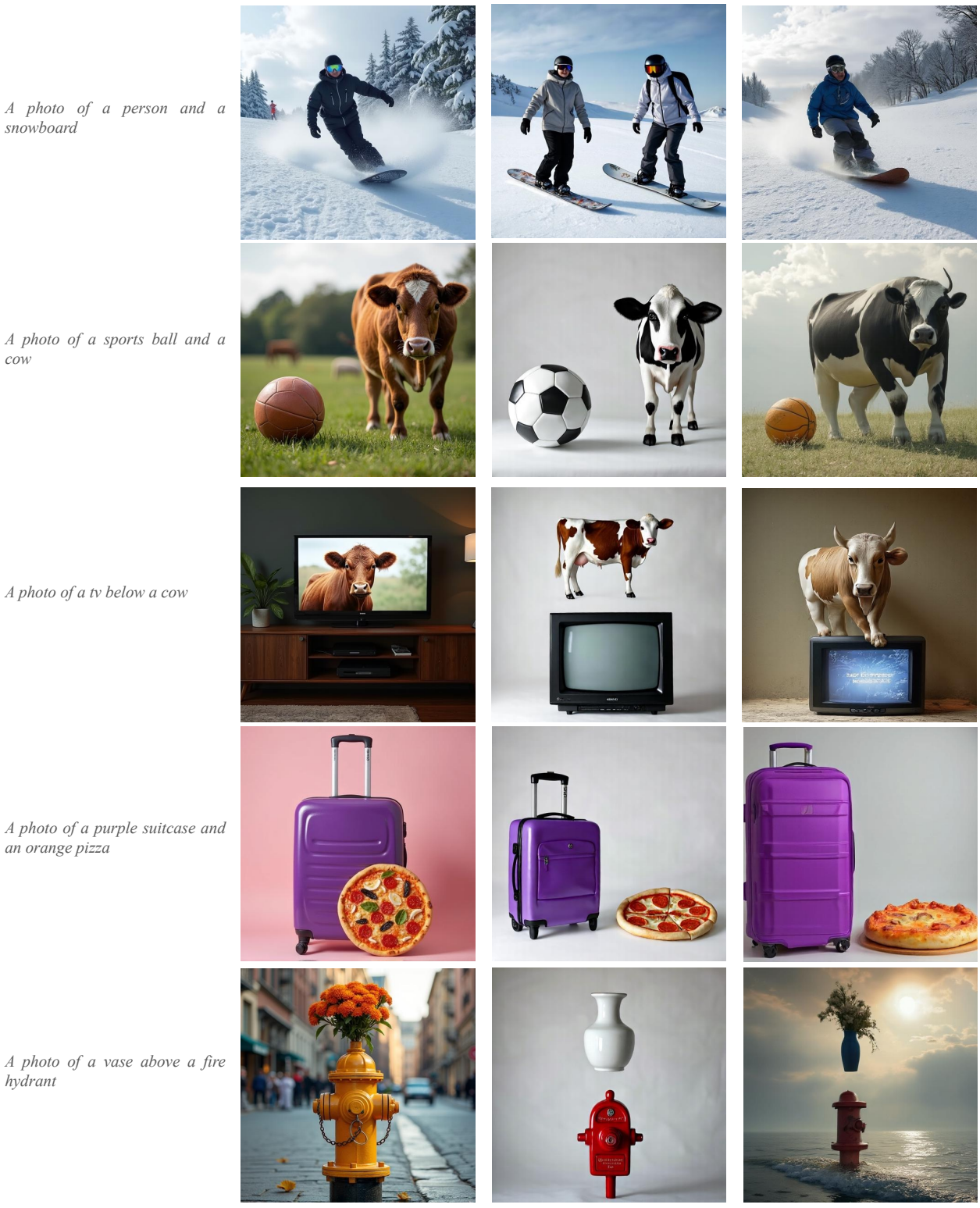


Figure xi. **More visualizations on GenEval**. We provide more comparisons between our method and Flow-GRPO when the evaluation reward is the same on the GenEval task.

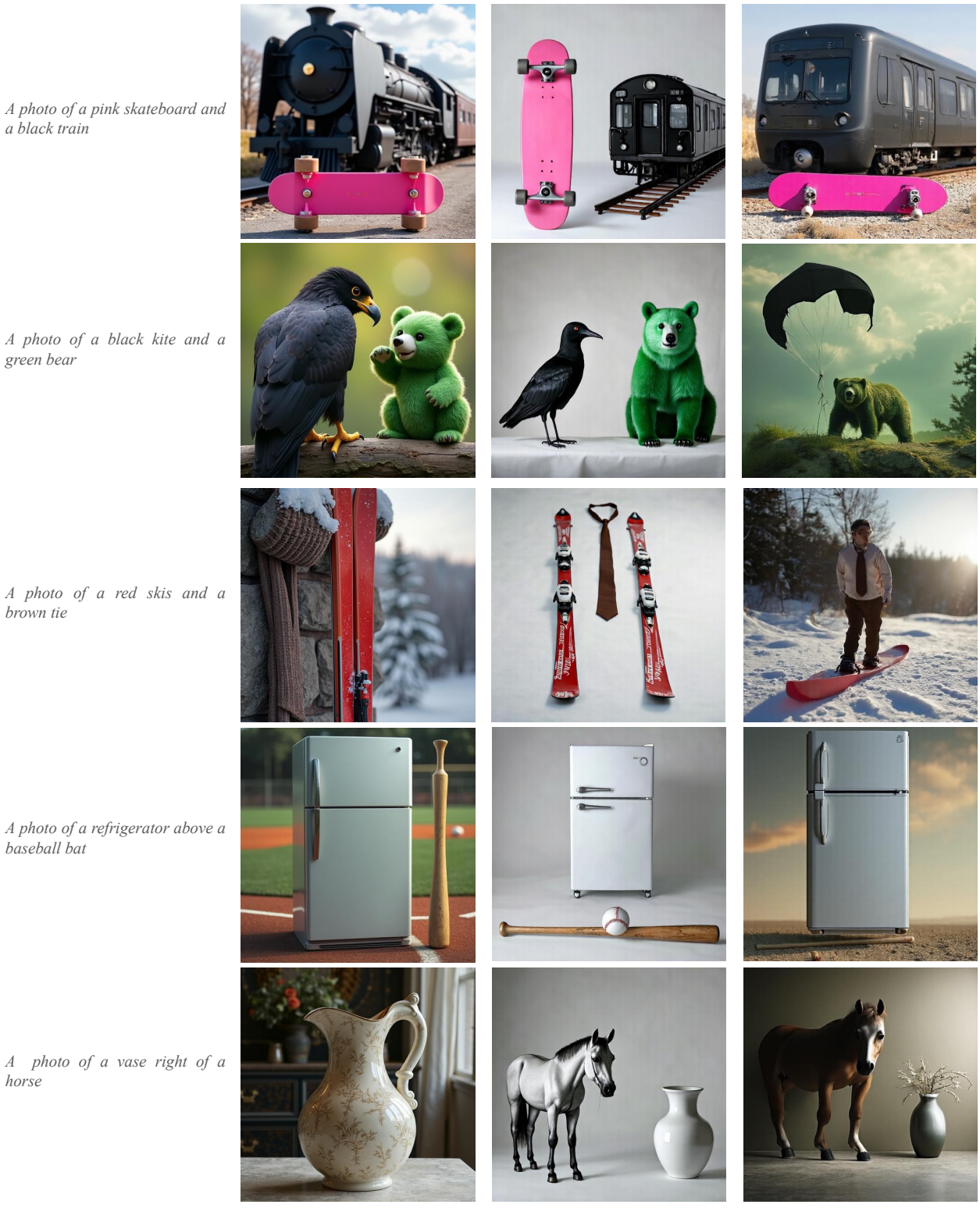


Figure xii. **More visualizations on GenEval**. We provide more comparisons between our method and Flow-GRPO when the evaluation reward is the same on the GenEval task.