

# 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:

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

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

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

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  $x_1, \dots, x_k$  and their corresponding explicit rewards  $r_1, \dots, r_k$ , the likelihood of the ranking  $x_1 \succ \dots \succ x_k$  is:

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

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

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

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:

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

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

$$\begin{aligned} \lim_{\tau \rightarrow 0} \mathcal{L}_{\text{GDRO}}^{k=2}(\theta) &= \sum_{i=1}^{2-1} \left( \log \sum_{m=i}^2 \exp(s_m) - \sum_{j=i}^2 q_i(j, \tau) s_j \right) \\ &= \log \sum_{m=1}^2 \exp(s_m) - \sum_{j=1}^2 q_1(j, \tau) s_j \\ &= \log(e^{s_1} + e^{s_2}) - s_1 \\ &= \log(e^{s_1}(1 + e^{s_2 - s_1})) - s_1 \\ &= \log(1 + e^{-\Delta}), \end{aligned} \quad (18)$$

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:

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

This means that, when  $k = 2, \tau \rightarrow 0$ , GDRO reduces to DPO. This is because when  $\tau \rightarrow 0$ , 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  $\beta$ .** 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. **Demonstration 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.

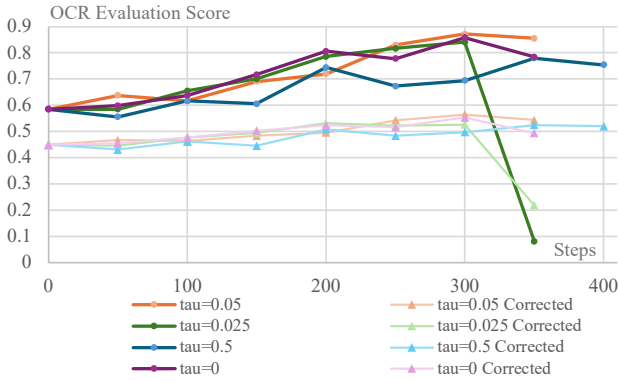


Figure viii. **Ablation study on temperature.** We provide the evaluation curves on different choices of the temperature  $\tau$ .

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  $\tau$ .** The temperature  $\tau$  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  $\tau$  is bigger, the distribution is more even and the differences between different reward values are reduced. When  $\tau$  is smaller, the differences of the reward values are amplified during calculation. We perform the ablation study on the choice of  $\tau$  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  $\tau$  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.

A medieval knight stands proudly, his shield emblazoned with the motto "defend the realm", reflecting the determination and honor of his cause. The shield's intricate design and weathered metal convey the battles and history it has witnessed.



A superhero stands proudly, his cape embroidered with "cape crusader" billowing in the wind. The scene is set at dusk, with a city skyline in the background, capturing the hero's determined silhouette.



A dragon's treasure chest, its lid emblazoned with "dragons 401k hands off", sits amidst a pile of gold and jewels, guarded by an ancient, slumbering dragon in a mystical, cavernous lair.



A realistic photograph of a zoo penguin exhibit, featuring a clear sign that reads "no flash photography again", surrounded by playful penguins and curious visitors.



A skyscraper window cleaner's bucket, tagged "wash away the grime", hangs from a high-rise in a bustling city, reflecting the sunlight and the towering buildings around it.



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.

A tennis racket with the tape intricately wrapped around the handle and strings, featuring the phrase " game set match " repeated in a stylish, bold font. the racket is set against a blurred tennis court background, emphasizing the dynamic sport.



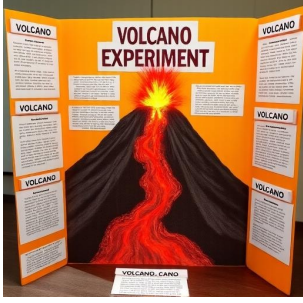
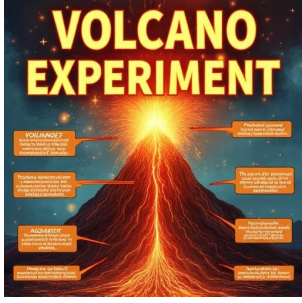
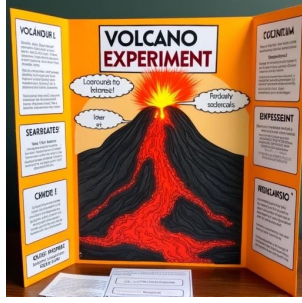
A vintage suitcase adorned with a " world traveler adventures " sticker, placed on a rustic wooden table. sunlight filters through a nearby window, casting warm, golden hues on the suitcase and highlighting the sticker's faded, nostalgic colors.



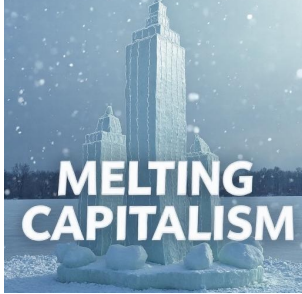
In a dimly lit antique shop, a vintage sign reads " vintage regret circa 1 9 9 9 ", hanging above a cluttered shelf filled with nostalgic items from the late 9 0 s, including old cds, vhs tapes, and retro toys.



A vibrant science fair poster titled " volcano experiment ", featuring a detailed illustration of a volcanic eruption with lava flowing down its slopes, surrounded by labeled diagrams and facts about volcanic processes, all set against a bright, engaging background.



A frozen lake hosts an ice carving festival, featuring an intricate ice sculpture titled " melting capitalism ". the sculpture depicts a melting skyscraper, with water cascading down its sides, symbolizing the erosion of capitalist structures. snowflakes gently fall around it, enhancing the serene yet powerful scene.



FLUX.1(0.58)

Flow-GRPO (0.87)

Ours (0.87)

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.

*A photo of a person and a snowboard*



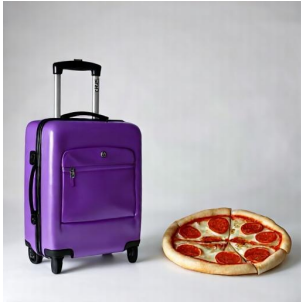
*A photo of a sports ball and a cow*



*A photo of a tv below a cow*



*A photo of a purple suitcase and an orange pizza*



*A photo of a vase above a fire hydrant*



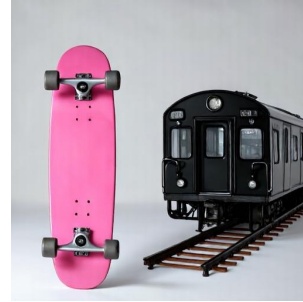
FLUX.1(0.58)

Flow-GRPO (0.85)

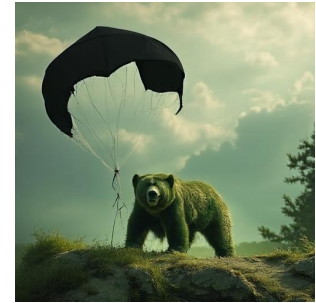
Ours (0.85)

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.

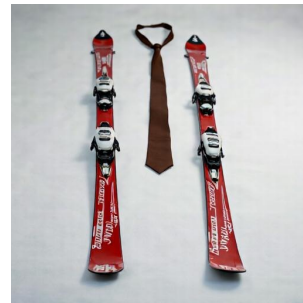
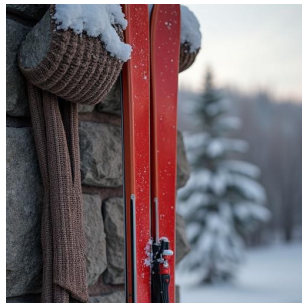
*A photo of a pink skateboard and a black train*



*A photo of a black kite and a green bear*



*A photo of a red skis and a brown tie*



*A photo of a refrigerator above a baseball bat*



*A photo of a vase right of a horse*



FLUX.1(0.58)

Flow-GRPO (0.85)

Ours (0.85)

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.