GRPO-Guard: Mitigating Implicit Over-Optimization in Flow Matching via Regulated Clipping

Jing Wang[1,2], Jiajun Liang[2*], Jie Liu[3], Henglin Liu[2,4], Gongye Liu[2,5], Jun Zheng[1], Wanyuan Pang[6], Ao Ma[7], Zhenyu Xie[1], Xintao Wang[2], Meng Wang[2], Pengfei Wan[2], Xiaodan Liang[1,8,9†]

[1]Shenzhen Campus of Sun Yat-Sen University, [2]Kling Team, Kuaishou Technology, [3]CUHK MMLab, [4]Tsinghua University, [5]HKUST
[6]USTB, [7]UCAS, [8]Peng Cheng Laboratory, [9]Guangdong Key Laboratory of Big Data Analysis and Processing

[*]Project Leader [†]Corresponding Author *Emails: wangj977@mail2.sysu.edu.cn, xdliang328@gmail.com*

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

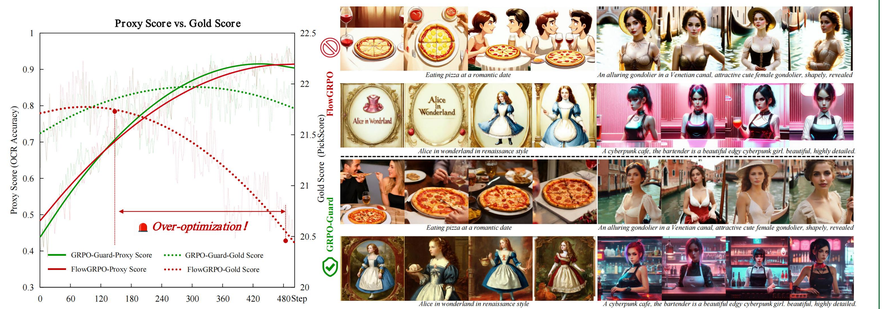


Figure 1. **Comparison between FlowGRPO and GRPO-Guard under over-optimization**. **Left**: The proxy score and gold score trends during training. As the proxy score increases, FlowGRPO rapidly enters an over-optimization phase, where the gold score continuously declines. **Right**: A visual comparison between FlowGRPO and GRPO-Guard. Due to severe reward hacking, FlowGRPO suffers from a drastic degradation in diversity, detail richness, visual quality, and text-image consistency (upper part). In contrast, GRPO-Guard maintains a stable gold score and high visual quality under a comparable proxy score, as shown in the bottom part of the figure.

In Diffusion GRPO, the policy model inevitably enters an implicit ***over-optimization stage***—while the proxy reward continues to increase, essential metrics such as image quality and text–prompt alignment deteriorate sharply, ultimately making the learned policy ***impractical for real-world use.***

Figure 9. Samples of the policy model at different training steps.

Method

GRPO

$$\mathcal{J}_{\text{policy}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\min\big(r_t^i(\theta)\,\hat{A}_t^i,\ \text{clip}(r_t^i(\theta),\,1-\epsilon,\,1+\epsilon)\,\hat{A}_t^i\big)\Big), \tag{5}$$

$$\begin{aligned}\log r_t(\theta) &= \log p_\theta(x_{t-1}|x_t,\boldsymbol{c}) - \log p_{\theta_{old}}(x_{t-1}|x_t,\boldsymbol{c})\\ &= -\frac{\|\mu_{\theta_{old}}(x_t,t)-\mu_\theta(x_t,t)+\sigma_t\sqrt{dt}\cdot\boldsymbol{\epsilon}\|^2}{2\sigma_t^2 dt}\\ &\quad+\frac{\|\mu_{\theta_{old}}(x_t,t)-\mu_{\theta_{old}}(x_t,t)+\sigma_t\sqrt{dt}\cdot\boldsymbol{\epsilon}\|^2}{2\sigma_t^2 dt}\\ &= -\frac{\|\Delta\mu_\theta+\sigma_t\sqrt{dt}\cdot\boldsymbol{\epsilon}\|^2}{2\sigma_t^2 dt}+\frac{\|\boldsymbol{\epsilon}\|^2}{2}\end{aligned}$$

Bias ➡ $$= \boxed{-\frac{\|\Delta\mu_\theta\|^2}{2\sigma_t^2 dt}} - \frac{\Delta\mu_\theta\cdot\boldsymbol{\epsilon}}{\sigma_t\sqrt{dt}} \tag{7}$$

RatioNorm ➡

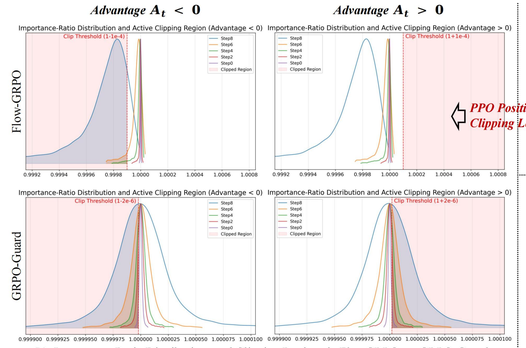
$$\begin{aligned}\log\hat{r}_t(\theta) &= \sigma_t\sqrt{dt}(\log r_t(\theta)+\frac{\|\Delta\mu_\theta\|^2}{2\sigma_t^2 dt})\\ &= -\Delta\mu_\theta\cdot\boldsymbol{\epsilon}\end{aligned} \tag{8}$$

$$\begin{aligned}\nabla_\theta\mathcal{J}(\theta) &= \sum_{t=0}^{T-1}\hat{A}_t\,\nabla_\theta\hat{r}_t(\theta) = \sum_{t=0}^{T-1}\hat{A}_t\,\hat{r}_t(\theta)\nabla_\theta\log\hat{r}_t(\theta)\\ &= \sum_{t=0}^{T-1}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}\left[\boldsymbol{\epsilon}\,\hat{A}_t\,\hat{r}_t(\theta)\,\nabla_\theta\mu_\theta(x_t|t)\right]\\ &= \sum_{t=0}^{T-1}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}\left[\boxed{\beta\,dt}\,\boldsymbol{\epsilon}\,\hat{A}_t\,\hat{r}_t(\theta)\nabla_\theta v_\theta(x_t|t)\right]\end{aligned} \tag{11}$$

$$\mathcal{J}_{\text{policy}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\delta\min\big(\hat{r_t^i}(\theta)\,\hat{A}_t^i,\ \text{clip}(\hat{r_t^i}(\theta),\,1-\epsilon,\,1+\epsilon)\,\hat{A}_t^i\big)\Big) \tag{12}$$

GRPO-Guard ➡

Analysis



(a) PPO Clip Mechanism

$r_t(\theta)$	A_t	Objective is Clipped	Gradient
$r_t(\theta)\in[1-\epsilon,\,1+\epsilon)$	+	No	√
$r_t(\theta)\in[1-\epsilon,\,1+\epsilon)$	−	No	√
$r_t(\theta)<1-\epsilon$	+	No	√
$r_t(\theta)<1-\epsilon$	−	Yes	×
$r_t(\theta)>1+\epsilon$	+	Yes	×
$r_t(\theta)>1+\epsilon$	−	No	√

$$L_{clip}^t(\theta) = \min(r_t(\theta)A_t, \text{clip}(r_t(\theta),1-\epsilon,1+\epsilon)A_t)$$

(b) Importance-Ratio Distributions and Clipping Regions in FlowGRPO vs. GRPO-Guard. **(c)** Mean and Standard Deviation of $r_t(\theta)$

Figure 2. **Comparison of $r(\theta)$ distributions between FlowGRPO and GRPO-Guard across timesteps.** (a) Ideally, the ratio distribution should have a mean near 1 and stable variance across timesteps to ensure effective clipping. (b) Under FlowGRPO, the distribution exhibits a leftward mean shift and increasing variance at low-noise timesteps, causing the clipping mechanism to fail—particularly for trajectories with positive advantages. In contrast, GRPO-Guard with RatioNorm preserves a balanced mean and consistent variance (c), enabling proper clipping and stable policy updates across all timesteps.

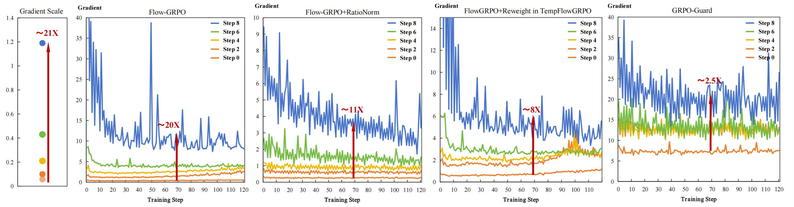


Figure 3. **Gradient magnitude differences across timesteps**. In FlowGRPO, gradient magnitudes vary by roughly 20× across timesteps, reflecting the large differences in gradient scale. GRPO-Guard substantially reduces this imbalance, limiting the variation to about 2.5× and preventing over-optimization under any single noise condition.

Setting	$\log r_t(\theta)$	Re-weight Scale	Gradient Scale
Baseline	$\log r_t(\theta)$	1	$\beta\frac{\Delta\mu_\theta+\sigma_t\sqrt{dt}\,\epsilon}{\sigma_t^2}$
Temp-Reweight [16]	$\log r_t(\theta)$	$\sigma_t\sqrt{dt}$	$\beta\frac{\sqrt{dt}\Delta\mu_\theta+\sigma_t dt\,\epsilon}{\sigma_t}$
Guard-Reweight	$\log r_t(\theta)$	$1/dt$	$\beta\frac{\Delta\mu_\theta+\sigma_t\sqrt{dt}\,\epsilon}{\sigma_t^2 dt}$
Mean-revised	$\log r_t(\theta)+\frac{\|\Delta\mu_\theta\|^2}{2\sigma_t^2 dt}$	1	$\beta\frac{\sqrt{dt}\,\epsilon}{\sigma_t}$
RatioNorm	$\sigma_t\sqrt{dt}(\log r_t(\theta)+\frac{\|\Delta\mu_\theta\|^2}{2\sigma_t^2 dt})$	1	$\beta\,dt\,\epsilon$
GRPO-Guard	$\sigma_t\sqrt{dt}(\log r_t(\theta)+\frac{\|\Delta\mu_\theta\|^2}{2\sigma_t^2 dt})$	$1/dt$	$\beta\,\epsilon$

Table 2. Ablation study on major components.

Ablation Study

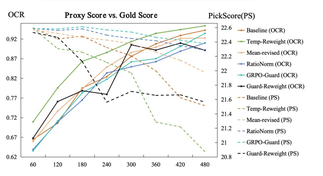


Figure 7. Training curves of the ablation study.

Experiments

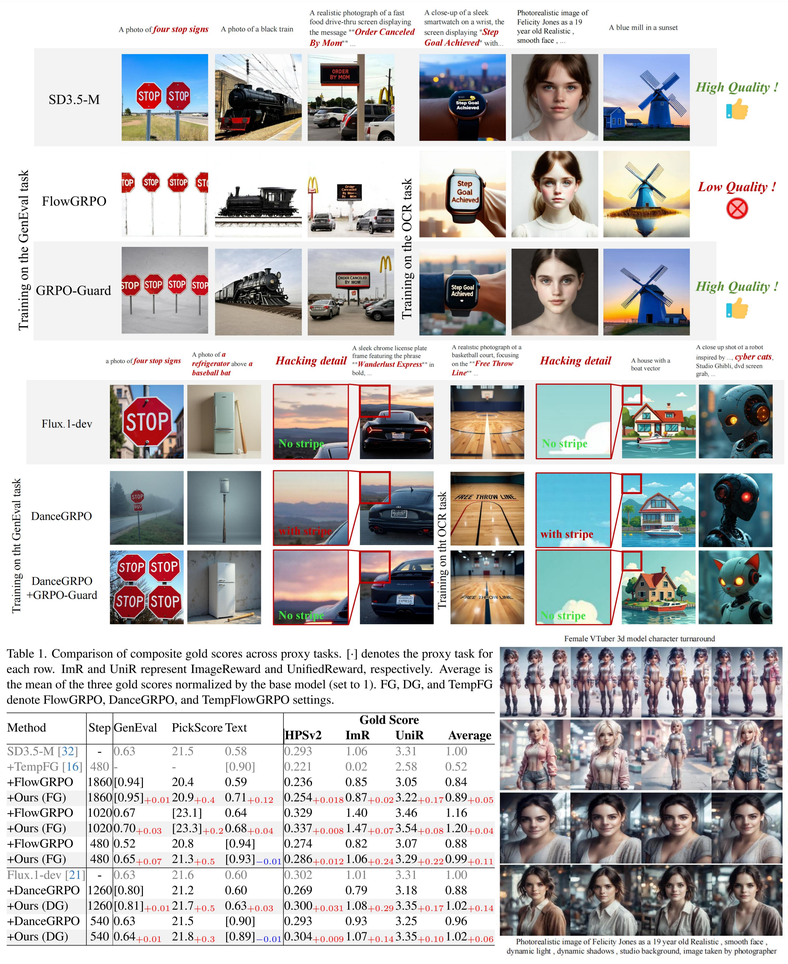


Table 1. Comparison of composite gold scores across proxy tasks. [·] denotes the proxy task for each row. ImR and UniR represent ImageReward and UnifiedReward, respectively. Average is the mean of the three gold scores normalized by the base model (set to 1). FG, DG, and TempFG denote FlowGRPO, DanceGRPO, and TempFlowGRPO settings.

Method	Step	GenEval	PickScore	Text	Gold Score HPSv2	Gold Score ImR	Gold Score UniR	Gold Score Average
SD3.5-M [32]	-	0.63	21.5	0.58	0.293	1.06	3.31	1.00
+TempFG [16]	480	-	-	[0.90]	0.221	0.02	2.58	0.52
+FlowGRPO	1860	[0.94]	20.4	0.59	0.236	0.85	3.05	0.84
+Ours (FG)	1860	[0.95]$_{+0.01}$	20.9$_{+0.4}$	0.71$_{+0.12}$	0.254$_{+0.018}$	0.87$_{+0.02}$	3.22$_{+0.17}$	0.89$_{+0.05}$
+FlowGRPO	1020	0.67	[23.1]	0.64	0.329	1.40	3.46	1.16
+Ours (FG)	1020	0.70$_{+0.03}$	[23.3]$_{+0.2}$	0.68$_{+0.04}$	0.337$_{+0.008}$	1.47$_{+0.07}$	3.54$_{+0.08}$	1.20$_{+0.04}$
+FlowGRPO	480	0.52	20.8	[0.94]	0.274	0.82	3.07	0.88
+Ours (FG)	480	0.65$_{+0.07}$	21.3$_{+0.5}$	[0.93]$_{-0.01}$	0.286$_{+0.012}$	1.06$_{+0.24}$	3.29$_{+0.22}$	0.99$_{+0.11}$
Flux.1-dev [21]	-	0.63	21.6	0.60	0.302	1.01	3.31	1.00
+DanceGRPO	1260	[0.80]	21.2	0.60	0.269	0.79	3.18	0.88
+Ours (DG)	1260	[0.81]$_{+0.01}$	21.7$_{+0.5}$	0.63$_{+0.03}$	0.300$_{+0.031}$	1.08$_{+0.29}$	3.35$_{+0.17}$	1.02$_{+0.14}$
+DanceGRPO	540	0.63	21.5	[0.90]	0.293	0.93	3.25	0.96
+Ours (DG)	540	0.64$_{+0.01}$	21.8$_{+0.3}$	[0.89]$_{-0.01}$	0.304$_{+0.009}$	1.07$_{+0.14}$	3.35$_{+0.10}$	1.02$_{+0.06}$