Reinforcing Diffusion Models by Direct Group Preference Optimization

Yihong Luo, Tianyang Hu, Jing Tang

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Background and Our Method

GRPO For Diffusion Models:

The GRPO relies on stochastic SDE for rollout and training, rather than the more efficient deterministic ODE. The policy-based method requires stochastic policy, however, the stochasticity in DM's policy comes from model agnostic Gaussian Noise.

Powerful but Slow in Training!

Diffusion-DPO:

The Diffusion-DPO does not relies on stochastic policy, therefor, its training is faster. However, it strictly relies on pairwise samples, which could not leverage rich relative information within groups like GRPO.

Fast but Less-Powerful!

DGPO (Direct Group Preference Optimization):

Key insight: the success of methods like GRPO stems from leveraging fine-grained, relative preference information, not from the policy-gradient formulation itself.

DGPO circumvents this problem by optimizing group-level preferences directly, *extending the DPO framework to handle* ***pairwise groups*** *instead of* ***pairwise samples***. This allows us to:

- Use Efficient Samplers
- Learn Directly from Preferences
- Train Efficiently

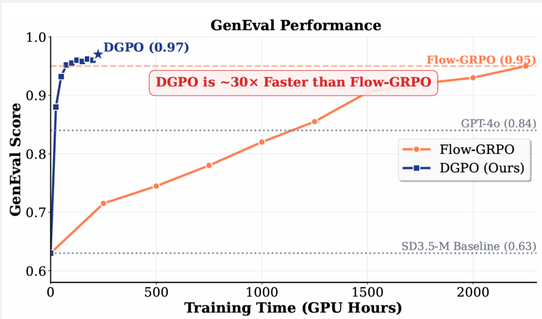
In particular, DGPO optimizes the following group-level reward.

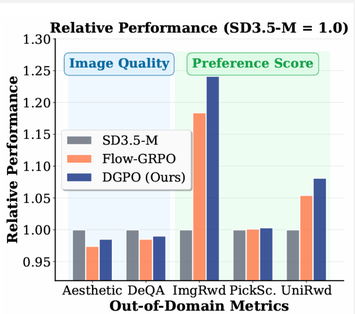
$$\max_\theta E_{(G^+,G^-,c)\sim D}\log p_\theta(G^+ \succ G^- \mid c) = E_{(G^+,G^-,c)\sim D}\log\sigma(R_\theta(G^+ \mid c) - R_\theta(G^- \mid c)).$$

Powerful and Fast in Training!

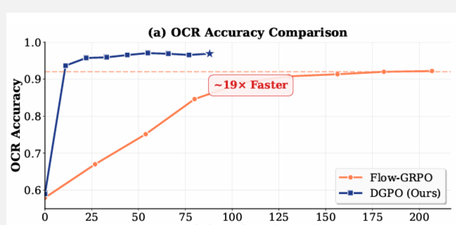
TL;DR, Pairwise group preference learning as powerful and efficient online RL native for Diffusion Models.

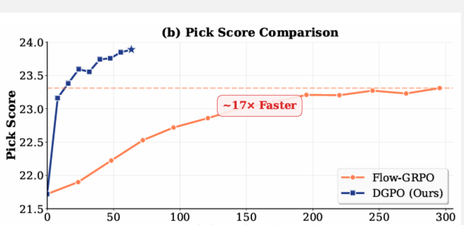
Empirical Evaluations





Ultra-Fast Training & Ultra-Powerful Performance





Model	Task Metric			Image Quality		Preference Score		
	GenEval	OCR Acc.	PickScore	Aesthetic	DeQA	ImgRwd	PickScore	UniRwd
SD3.5-M	0.63	0.59	21.72	5.39	4.07	0.87	22.34	3.33
Compositional Image Generation:								
Flow-GRPO	0.95	—	—	5.25	4.01	1.03	22.37	3.51
DGPO (Ours)	0.97	—	—	5.31	4.03	1.08	22.41	3.60
Visual Text Rendering:								
Flow-GRPO	—	0.92	—	5.32	4.06	0.95	22.44	3.42
DGPO (Ours)	—	0.96	—	5.37	4.09	1.02	22.52	3.48
Human Preference Alignment:								
Flow-GRPO	—	—	23.31	5.92	4.22	1.28	23.53	3.66
DGPO (Ours)	—	—	23.89	6.08	4.40	1.32	23.91	3.74

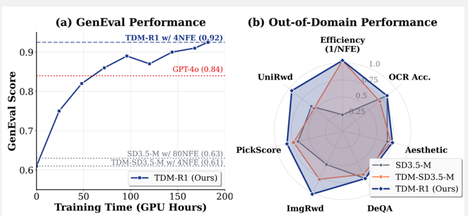
Paper

Code

Consistent improvement over the Flow-GRPO.

Our Follow-up Work in Few-Step Model RL

TDM-R1:



Paper

Code

TDM-R1 is the first systematical attempt on leveraging **Non-Differentiable Reward for reinforcing few-step models** with significant improvements, where training with a single reward signal yields consistent OOD improvements across different backbones

TDM-R1 decouples the learning process into surrogate reward learning and generator learning.

- The surrogate reward learning is inspired by *DGPO*, which learns step-wise reward estimation along the student trajectory.
- The generator is then learned with standard RLHF objective.

4NFE-Samples

Model	NFE	Verifiable Metric		Image Quality		Preference Score		
		GenEval	OCR Acc.	Aesthetic	DeQA	ImgRwd	PickScore	UniRwd
SD3.5-M (Esser et al., 2024)	80	0.63	0.59	5.39	4.07	0.87	22.34	3.33
Flow-GRPO† (Liu et al., 2025)	80	0.95	—	5.25	4.01	1.03	22.37	3.51
DGPO† (Luo et al., 2025d)	80	0.97	—	5.31	4.03	1.08	22.41	3.60
Flow-GRPO† (Liu et al., 2025)	80	—	0.92	5.32	4.06	0.95	22.44	3.42
DGPO† (Luo et al., 2025d)	80	—	0.96	5.37	4.09	1.02	22.52	3.48
TDM-SD3.5-M (Luo et al., 2025c)	4	0.61	0.55	5.41	4.05	0.99	22.36	3.30
Compositional Image Generation:								
p_θ in Surrogate Reward	80	0.89	0.52	5.35	4.01	0.94	22.31	3.48
TDM-R1 (Ours)	4	0.92	0.59	5.42	4.07	1.11	22.39	3.55
Visual Text Rendering:								
p_θ in Surrogate Reward	80	0.64	0.92	5.41	4.05	0.98	22.37	3.43
TDM-R1 (Ours)	4	0.67	0.95	5.45	4.11	1.09	22.62	3.51

Highly powerful RL algorithm native for Few-step generative models!