DanceGRPO: Unleashing GRPO on Visual Generation

Speaker: Zeyue Xue

The University of Hong Kong

30/06/2025

Theoretical Framework

Forward SDE for diffusion model: $\mathrm{d}\mathbf{z}_t = f_t \mathbf{z}_t \mathrm{d}t + g_t \mathrm{d}\mathbf{w}$

Backward SDE for diffusion model: $\mathrm{d}\mathbf{z}_t = \left(f_t \mathbf{z}_t - \frac{1+\varepsilon_t^2}{2} g_t^2 \nabla \log p_t(\mathbf{z_t}) \right) \mathrm{d}t + \varepsilon_t g_t \mathrm{d}\mathbf{w}$

However, the forward process of rectified flow is defined by an ODE: $\mathrm{d}\mathbf{z}_t = \mathbf{u}_t \mathrm{d}t$

Motivated by stochastic interpolants,

we give an SDE case for rectified flows: $\mathrm{d}\mathbf{z}_t = (\mathbf{u}_t - \frac{1}{2}\varepsilon_t^2 \nabla \log p_t(\mathbf{z}_t))\mathrm{d}t + \varepsilon_t \mathrm{d}\mathbf{w}$

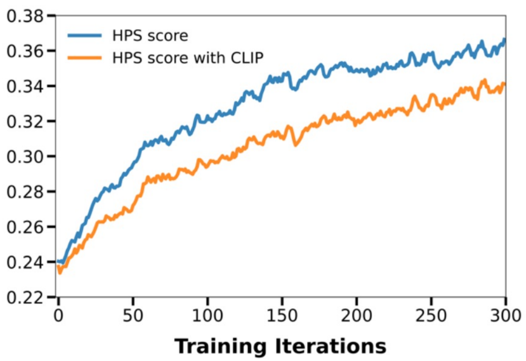
DanceGRPO

Algorithm 1 DanceGRPO Training Algorithm

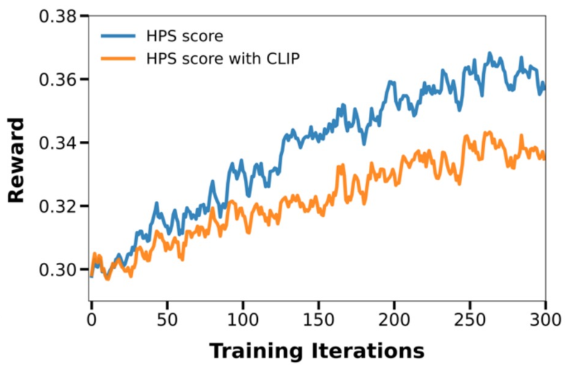
Require: Initial policy model π_θ; reward models $\{R_k\}_{k=1}^{K}$; prompt dataset $\mathcal{D}$; timestep selection ratio τ; total sampling steps T
Ensure: Optimized policy model π_θ

```
for training iteration = 1 to M do
    Sample batch D_b ~ D                                        ▷ Batch of prompts
    Update old policy: π_θold ← π_θ
    for each prompt c ∈ D_b do
        Generate G samples: {o_i}_{i=1}^G ~ π_θold(·|c) with the same random initialization noise
        Compute rewards {r_i^k}_{i=1}^G using each R_k
        for each sample i ∈ 1..G do
            Calculate multi-reward advantage: A_i ← Σ_{k=1}^K (r_i^k − μ^k)/σ^k    ▷ μ^k, σ^k per-reward statistics
        end for
        Subsample ⌈τT⌉ timesteps T_sub ⊂ {1..T}
        for t ∈ T_sub do
            Update policy via gradient ascent: θ ← θ + η∇_θ J
        end for
    end for
end for
```

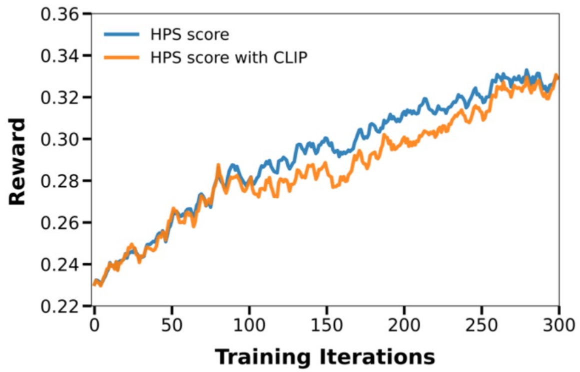
Results



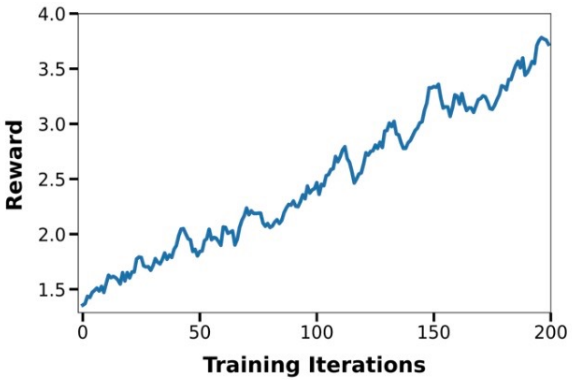
(a) Reward on aesthetics of Stable Diffusion



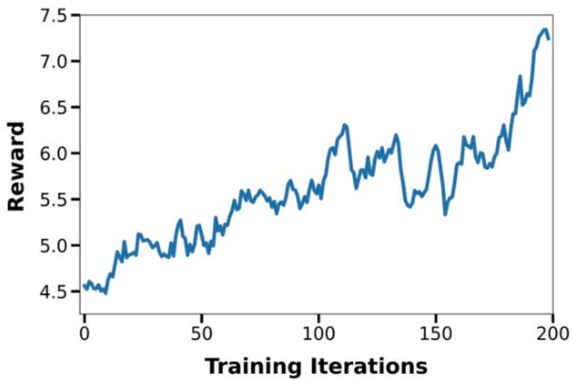
(b) Reward on aesthetics of FLUX.1-dev



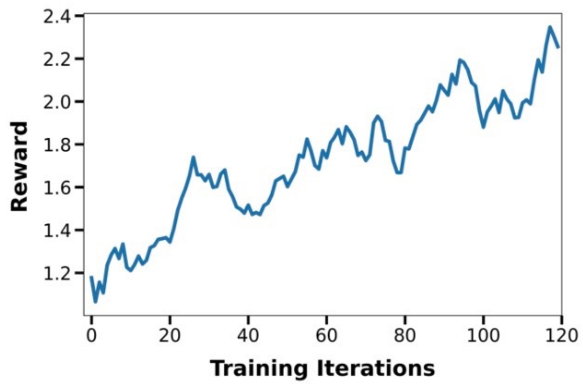
(c) Reward on aesthetics of HunyuanVideo-T2I



(a) Reward on Videoalign motion quality (text-to-video)



(b) Reward on Videoalign visual aesthetics quality (text-to-video)



(c) Reward on Videoalign motion quality (image-to-video)

Results

Models	HPS-v2.1 [19]	CLIP Score [20]	Pick-a-Pic [33]	GenEval [21]
Stable Diffusion	0.239	0.363	0.202	0.421
Stable Diffusion with HPS-v2.1	**0.365**	0.380	**0.217**	0.521
Stable Diffusion with HPS-v2.1&CLIP Score	0.335	**0.395**	0.215	**0.522**

Models	HPS-v2.1 [21]	CLIP Score [22]	Pick-a-Pic [35]	GenEval [23]
FLUX	0.304	0.405	0.224	0.659
FLUX with HPS-v2.1	**0.372**	0.376	**0.230**	0.561
FLUX with HPS-v2.1&CLIP Score	0.343	**0.427**	0.228	**0.687**

Visualization

Prompt: A man lying down on green grass, gazing at the stars during an evening at a countryside villa

Prompt: A sinister man with red eyes speaking, very close shot, Cthulhu

Prompt: A chubby baby playing with toys in the snow

Prompt: Generate a picture of a blue sports car parked on the road, metal texture

Open Questions

1. How can we speed up training while maintaining quality?
2. Can rule-based rewards work for visual generation?
3. What's the best reward model design for RL, CLIP, VLM, AI feedback or mix?
4. How can we improve algorithms, such as following DAPO

Open Questions

5. Downstream applications (medical, 3D, personalization, editing, etc.).

6. Should video reward models analyze every frame?

7. How can we design a joint GRPO algorithm for LLM and diffusion/flow?

8. How can we avoid reward hacking, other than model merging/mixing/EMA.

Q&A

Thanks!