ICML
International Conference
On Machine Learning

Flash-GRPO: Efficient Alignment for Video Diffusion via One-Step Policy Optimization

Authors: Xiaoxuan He, Siming Fu, Zeyue Xue, Weijie Wang, Ruizhe He, Yuming Li, Dacheng Yin, Shuai Dong, Haoyang Huang, Hongfa Wang, Nan Duan, Bohan Zhuang

ICML 2026 Presentation

Motivation

Aligning a 14B parameter video model typically demands hundreds of GPU days per experiment, imposing a scalability bottleneck that restricts both research iteration and practical deployment.

Overview

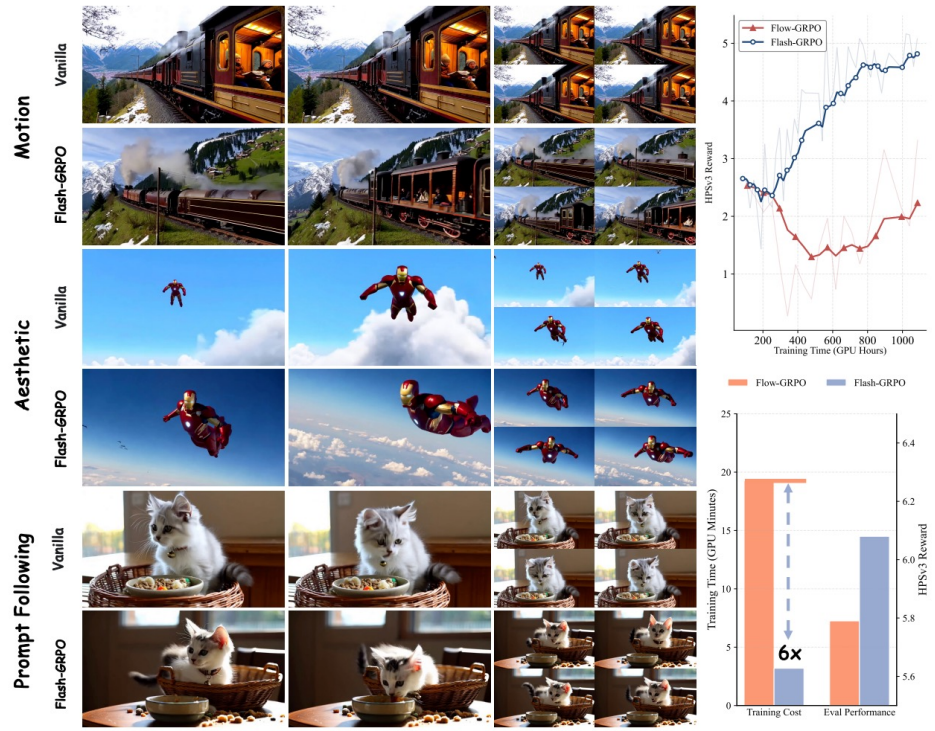

Figure 1 Overview of Flash-GRPO performance. **(Left)** Qualitative comparison across three dimensions: Motion, Aesthetic, and Prompt Following. Flash-GRPO generates videos with enhanced temporal dynamics (train sequence), improved visual quality (Iron Man), and better prompt adherence (cat with food bowl). **(Top Right)** Training reward curves showing that Flash-GRPO achieves stable monotonic improvement while Flow-GRPO exhibits slower convergence in training time. **(Bottom Right)** Efficiency comparison: Flash-GRPO achieves 6× acceleration in training cost while attaining higher evaluation performance.

Method

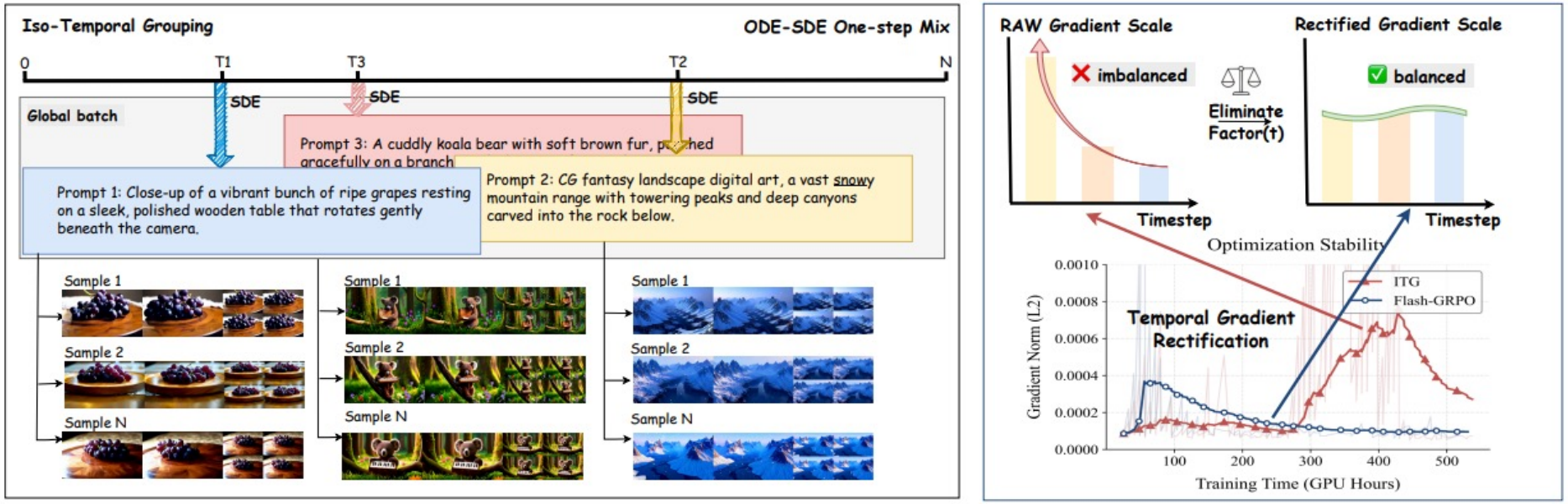

Figure 2 Overview of the Flash-GRPO Framework. **(Left)** Iso-temporal Grouping: each prompt performs ODE-to-SDE transition at a single sampled timestep for exploration and gradient computation, while other timesteps use deterministic ODE for accurate reward signals. Rollouts within each group share this transition timestep but differ in initial noise, factorizing policy-induced variance from timestep-induced variance. **(Right)** Temporal Gradient Rectification: the SDE discretization introduces a time-dependent scaling factor $\lambda(t)$ that causes gradient magnitudes to vary by orders of magnitude. Normalizing by $1/\lambda(t)$ ensures uniform contribution across timesteps, eliminating discretization-induced optimization bias.

Analysis

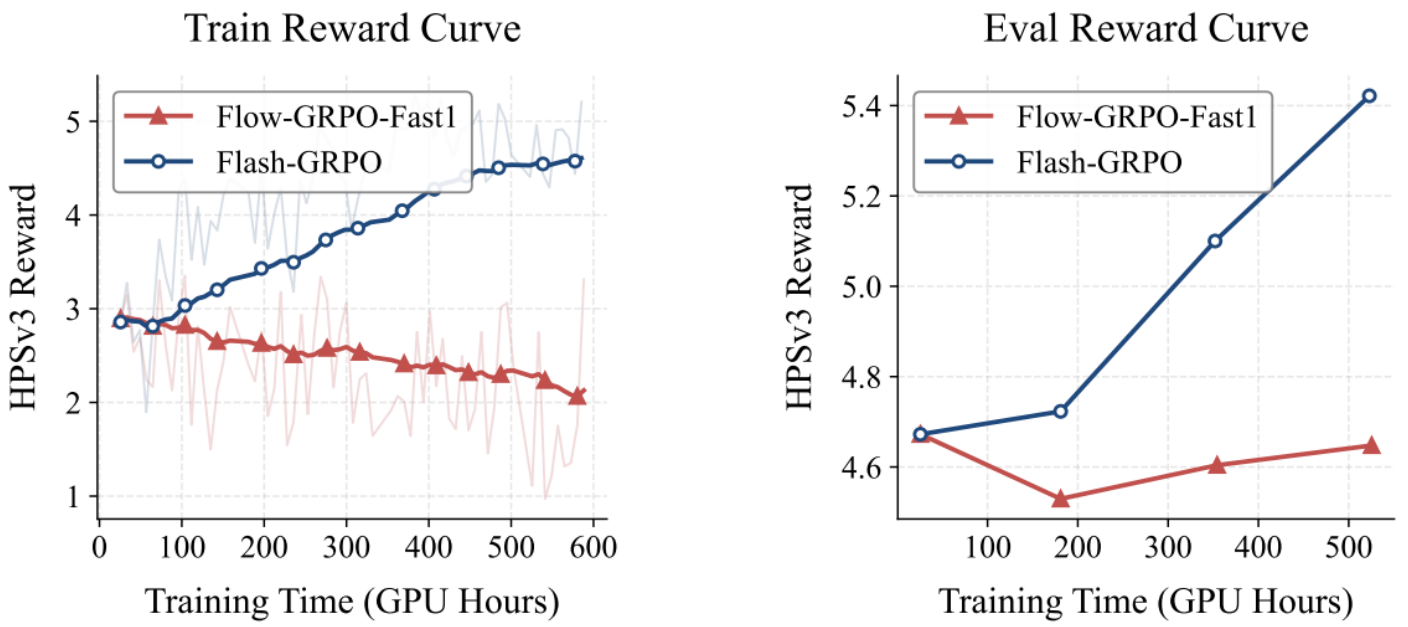

Figure 4 HPSv3 reward curves. Flow-GRPO-Fast1 suffers from optimization collapse on both training (Left) and evaluation (Right), while Flash-GRPO maintains stable convergence.

Results

Table 1 Detailed comparison of **General Video Quality** using VBench metrics. We evaluate aesthetic quality, image quality, subject consistency, and object class to ensure the RL fine-tuning retains the generative capability of the backbone model. We reproduce the official VBench results; $*$ indicates our own reproduction results (mismatch). Best scores are in blue.

Method	GPU Hours	Aesthetic Quality ↑	Imaging Quality ↑	Subject Consistency ↑	Object Class ↑
CogVideoX-2B [29]	–	61.07	62.37	96.52	86.48
Hunyuan-Video [9]	–	60.36	67.56	97.37	86.10
Wan2.1-T2V-1.3B [26]	–	65.46	66.79*/67.01	97.56	88.84*/88.81
Flow-GRPO-Fast1	350	65.92	65.96	98.46	88.15
Flow-GRPO	350	65.79	68.60	97.28	87.92
Flash-GRPO	350	66.43	68.28	98.70	90.00

Results

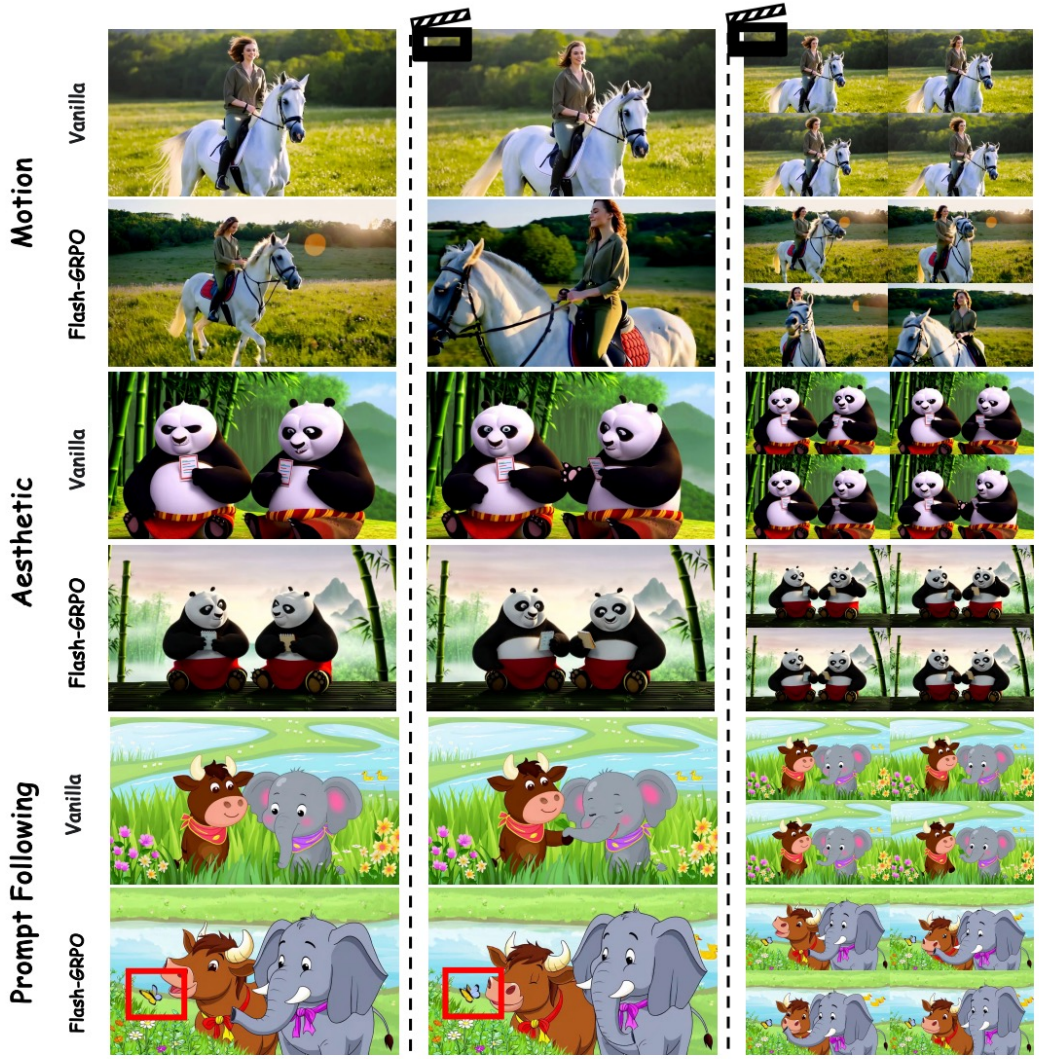

Figure 3 Qualitative comparison between vanilla Wan2.1 (odd rows) and Flash-GRPO (even rows) across three dimensions: Motion, Aesthetic, and Prompt Following. Flash-GRPO produces videos with enhanced temporal dynamics (horse riding sequence), improved visual quality and richer details (panda scene), and better prompt adherence with additional elements (cartoon animals with butterfly, highlighted in red boxes).

THANKS