MANIFOLD-AWARE EXPLORATION FOR REINFORCEMENT LEARNING IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) methods for video generation like FlowGRPO remain far less reliable than their counterparts for language models and images. This gap arises because video generation has a complex solution space, and the ODE-to-SDE conversion used for exploration can inject excess noise, lowering rollout quality and making reward estimates less reliable, which destabilizes post-training alignment. To address this problem, we view the pre-trained model as defining a valid video data manifold and formulate the core problem as constraining exploration within the vicinity of this manifold, ensuring that rollout quality is preserved and reward estimates remain reliable. We propose **SAGE-GRPO** (Stable Alignment via Guided Exploration), which applies constraints at both micro and macro levels. At the micro level, we derive a *precise manifold-aware SDE* with a logarithmic curvature correction and introduce a *gradient norm equalizer* to stabilize sampling and updates across timesteps. At the macro level, we use a *dual trust region* with a periodic moving anchor and stepwise constraints so that the trust region tracks checkpoints that are closer to the manifold and limits long-horizon drift. We evaluate SAGE-GRPO on HunyuanVideo1.5 using VideoAlign and HPSV3 as reward models, and generalize to images on SD3.5-medium and FLUX.1-dev with HPSV3. We observe consistent gains in VQ, MQ, TA, and visual metrics, demonstrating superior performance in reward maximization and overall generation quality. Our code will be released to the public.

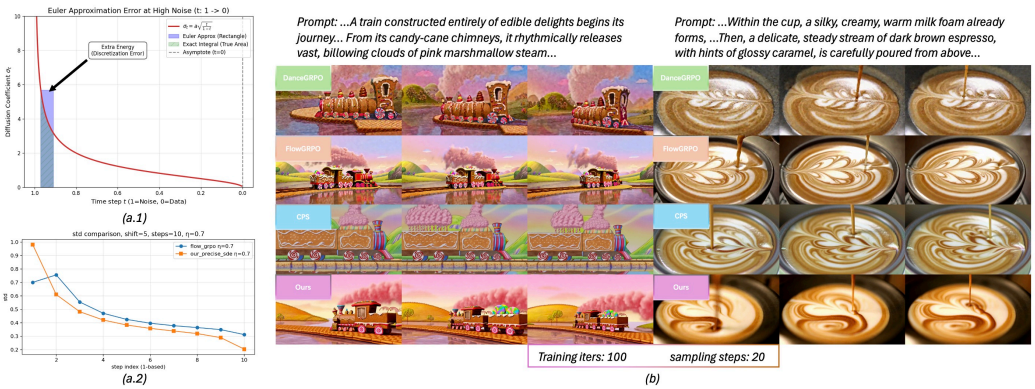


Figure 1: **Illustration of SAGE-GRPO. (Left)** (a.1) At a higher noise region, Euler-style discretization introduces a purple region of extra energy (discretization error) beyond the true integral; we focus on the true integral region below, not this extra energy. (a.2) Our precise SDE removes unnecessary noise energy in high-noise regions, enabling more precise exploration and a better-learned data manifold. **(Right)** (b) Our method with improved exploration yields more stable and better-aligned generations compared with DanceGRPO, FlowGRPO, and CPS.

1 INTRODUCTION

Group Relative Policy Optimization (GRPO) is a direct way to align video generation models with reward signals (Ho et al., 2020; Song et al., 2020b;a; Ma et al., 2025a; Kong et al., 2024; Wu

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

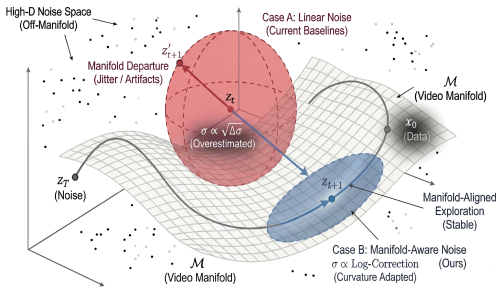


Figure 2: **Geometric interpretation of noise injection strategies.** Conventional linear SDEs (red) use first-order approximations, causing off-manifold drift and temporal jitter. Our Manifold-Aware SDE (blue) uses a logarithmic correction to guide exploration towards the manifold trajectory.

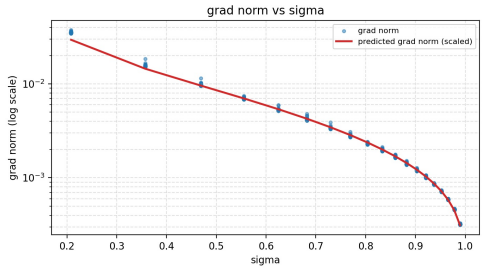


Figure 3: Empirical gradient norm imbalance across noise levels. Observed norms (blue) decrease rapidly as σ increases and match the predicted relationship (red) $\|\nabla \log \pi\| \propto 1/\Sigma_t^{1/2}$, leading to vanishing gradients at high noise ($\sigma \to 1$) and exploding gradients at low noise ($\sigma \to 0$).

et al., 2025; Wan et al., 2025; Gao et al., 2025), but it has not yet been as reliable for video as it is for language models and images (Guo et al., 2025; Shao et al., 2024; Achiam et al., 2023; Shen et al., 2025). We focus on the dominant family of *flow-matching* Diffusion-Transformer (DiT) video generators, exemplified by HunyuanVideo1.5 (Wu et al., 2025), Wan (Wan et al., 2025), and Seedance (Gao et al., 2025), where rollouts are produced by integrating an ODE backward from noise to video. Within this class, applying GRPO requires drawing a group of rollouts by converting the deterministic ODE sampler into an SDE sampler so that the policy can explore through diverse samples (Li et al., 2025a). Video generation has a large, structured solution space, so this exploration is easily disturbed. Current GRPO baselines for video generation such as DanceGRPO and FlowGRPO rely on an Euler-style discretization and first-order approximations when deriving the SDE noise standard deviation (as shown in Table 1, details in Section A) (Black et al., 2023; Liu et al., 2025b; Xue et al., 2025). The resulting first-order truncation error can inject excess noise energy during sampling (shown in Figure 1(a.1)), which lowers rollout quality in high-noise steps and makes reward evaluation less reliable. This raises the following question: *how can we obtain an accurate sampling path that improves rollout quality and stabilizes GRPO for video generation?*

In on-policy exploration, we start from an initial pre-trained flow-matching model (Liu et al., 2022; Lipman et al., 2022; Wang et al., 2024), which we treat as defining a valid data manifold $\mathcal{M} \subset \mathbb{R}^D$. Since the pre-trained model still exhibits flaws, we design RL algorithms like GRPO to improve the performance of the policy model via updating the parameters θ with exploration; however, we treat the reward as a guiding signal rather than the final target, as the true optimization objective is the real video data manifold $\mathcal{M}$. Consequently, the exploration must keep trajectories near $\mathcal{M}$ so that rollouts remain valid. As shown in Figure 2, previous style SDE exploration can overestimate the noise variance (red), push z_t away from $\mathcal{M}$, and produce temporal jitter. We therefore define the core problem of GRPO for video generation as **how to constrain exploration within the vicinity of the data manifold** so that each update improves rollouts while keeping reward evaluation reliable.

We propose **SAGE-GRPO** (Stable Alignment via Guided Exploration), which organizes exploration at both micro and macro levels around the manifold. At the *micro level*, we refine the discrete SDE and couple it with a *gradient normalization equalizer* as part of micro-scale exploration. Concretely, instead of using an area-based first-order variance approximation, we compute the noise variance by integrating the diffusion coefficient over each step and add a logarithmic correction $\log\left(\frac{1-\sigma_{t+\Delta t}}{1-\sigma_t}\right)$ that captures the curvature of the signal-decay coefficient $(1-\sigma_t)$, yielding a more accurate variance for ODE-to-SDE exploration. As in Figure 1(a.1), this corresponds to integrating only the effective energy under the curve rather than the extra discretization area. Correspondingly, Figure 1(a.2) shows that the resulting precise SDE uses smaller variance while staying closer to the underlying video manifold, and Figure 1(b) demonstrates how this directly translates into improved visual quality. Even with this corrected SDE, the diffusion process still has an inherent signal-to-noise imbalance across timesteps: gradients vanish at high noise ($t \to 1$) and explode at low noise ($t \to 0$), which biases learning toward certain phases. The gradient normalization equalizer rescales the per-timestep gradient magnitude so that all timesteps contribute updates of comparable scale to the policy, which makes micro-level exploration more precise and stable.

Table 1: Comparison of SDE noise injection strategies used in video GRPO.

Method	**Standard Deviation** $\Sigma_t^{1/2}$
DanceGRPO	$\eta\sqrt{\sigma_t - \sigma_{t+1}}$
FlowGRPO	$\eta\sqrt{\frac{\sigma_t}{1-\sigma_t}(\sigma_t - \sigma_{t+1})}$
Ours	$\eta\sqrt{\left[-(\sigma_t - \sigma_{t+1}) + \ln\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right)\right]}$

With precise micro-level exploration, the policy after N steps updates tends to move closer to the data manifold; periodically updating a reference model from this trajectory therefore creates a trust region centered at a more manifold-consistent policy. This reduces long-horizon drift and helps avoid off-manifold local optima, as suggested by the red region in Figure 2. Traditional Fixed KL constraints $D_{KL}(\pi_\theta||\pi_0)$ anchor the policy to the initial model π_0, but as training progresses the optimal policy π^* may be far from π_0, which causes underfitting. Step-wise KL constraints $D_{KL}(\pi_\theta||\pi_{k-1})$ limit the magnitude of parameter updates per step (velocity control), ensuring smooth local transitions, but they only constrain the instantaneous update direction ∇_θ and do not bound the cumulative displacement $\|\theta_k - \theta_0\|$ from the initial parameters. This allows unbounded drift: even if each step is small, the policy can move slowly but consistently away from the manifold over many steps, eventually leading to degradation or reward hacking. To counteract drift while preserving plasticity, we introduce a *Periodical Moving Anchor* that updates the reference policy π_{ref} every N steps, creating a dynamic trust region that repeatedly recenters exploration near a manifold-consistent policy. We combine the moving anchor with step-wise constraints into a *Dual Trust Region* objective that provides position control towards the manifold and velocity control between successive policies, forming a position-velocity controller that enables sustained plasticity.

We evaluate **SAGE-GRPO** on HunyuanVideo1.5 (Wu et al., 2025) using VideoAlign (Liu et al., 2025c) and HPSV3 as reward models, and we observe consistent gains over baselines such as DanceGRPO (Xue et al., 2025), FlowGRPO (Liu et al., 2025b), and CPS (Wang & Yu, 2025) across reward maximization, temporal fidelity, and visual metrics. To further demonstrate generalizability, we also evaluate SAGE-GRPO on image generation using SD3.5-medium and FLUX.1-dev with HPSV3, yielding better performance on both reward scores and final image quality. Extensive ablations confirm that both the micro-level precise exploration design and the macro-level dual trust region mechanism are necessary to reduce the stability–plasticity gap.

Our main contributions are as follows:

- We formulate GRPO for video generation as a manifold-constrained exploration problem and show that the ODE-to-SDE conversions used in existing methods can inject excess noise in high-noise steps, which reduces rollout quality and makes reward-guided updates less reliable.
- At the micro-level, we constrain precise exploration via a *Manifold-Aware SDE* for manifold-consistent sampling and a *Gradient Norm Equalizer* for balanced timestep updates.
- At the macro-level, we propose a *Dual Trust Region* with moving anchors and step-wise constraints to ensure manifold-consistent tracking and prevent drift during long-horizon exploration.

2 Related Work

Reinforcement Learning for Diffusion and Flow Matching Models. Reinforcement learning has been adapted to fine-tune diffusion and flow matching models (Liu et al., 2025b; Xue et al., 2025; Xu et al., 2023; Jiang et al., 2025; Wallace et al., 2024; Xu et al., 2025; Lan et al., 2025; Jin et al., 2025; Zhang et al., 2026) for alignment with human preferences. Early approaches such as DDPO (Black et al., 2023) and DPOK (Fan et al., 2023) treated the denoising process as a Markov Decision Process to enable policy gradient estimation. Inspired by GRPO in language models (Shao et al., 2024; Guo et al., 2025), *FlowGRPO* (Liu et al., 2025b) and *DanceGRPO* (Xue et al., 2025) adapted GRPO to visual generation via ODE-to-SDE conversion for stochastic exploration (Li et al., 2025a). Further exploring this space, *Diffusion-NFT* (Zheng et al., 2025) avoids the reverse SDE entirely, building contrastive supervised learning on the forward process, and *TempFlow-GRPO* (He et al., 2025b) stays within reverse-process GRPO with a temporal flow weighting scheme. However, existing approaches

use first-order approximations that push exploration off the data manifold and fail to address gradient imbalance across timesteps.

Preference Alignment for Video Generation. Aligning video generation models with human preferences is an active research area (Zheng et al., 2024; Long et al., 2025; Huang et al., 2024; Lu et al., 2025; He et al., 2025a). Building on video diffusion models (Wan et al., 2025; Kong et al., 2024; Gao et al., 2025), researchers have developed video reward models (Liu et al., 2025c; Xu et al., 2024; Mi et al., 2025; Zhang et al., 2025) and alignment algorithms (Li et al., 2024; Gambashidze et al., 2024; Yu et al., 2024; Zhou et al., 2025; Jia et al., 2025). DanceGRPO (Xue et al., 2025) extends image-based RL to video, while Self-paced GRPO (Li et al., 2025b) proposes curriculum learning that dynamically adjusts reward weights. However, current alignment frameworks face a stability-plasticity dilemma: strict constraints (e.g., fixed KL anchored to initialization) limit plasticity, while relaxed constraints trigger reward hacking or catastrophic forgetting (Liu et al., 2025a; Li et al., 2025c). Unlike existing approaches that rely on heuristic scheduling or static anchors, our method integrates manifold-aware dynamics with a dual trust region to resolve this tension.

3 METHODOLOGY

We formulate the problem of video alignment as maximizing the expected reward $J(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim \pi_\theta}[R(\mathbf{x}_0)]$ within a Group Relative Policy Optimization (GRPO) framework. However, a standard application of GRPO to video diffusion models faces specific challenges in maintaining stable and effective exploration on the video manifold. **SAGE-GRPO** addresses these challenges by designing a unified exploration strategy that operates from micro-level noise injection to macro-level policy constraints so that every exploration step remains valid and balanced across timesteps.

3.1 PRELIMINARIES: FLOW MATCHING AND GROUP RELATIVE POLICY OPTIMIZATION

Flow Matching and Rectified Flow. Flow Matching models generation as transport along a probability path $p_t(\mathbf{x})$ via an ordinary differential equation (ODE):

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t), \tag{1}$$

where $\mathbf{v}_\theta$ is a neural velocity field. Rectified Flow uses the linear interpolation path:

$$\mathbf{x}_t = (1 - \sigma_t)\mathbf{x}_0 + \sigma_t \mathbf{z}_1, \tag{2}$$

which implies the velocity field:

$$\mathbf{v}_\theta(\mathbf{x}_t, t) = \frac{d\mathbf{x}_t}{dt} = -\frac{d\sigma_t}{dt}(\mathbf{x}_0 - \mathbf{z}_1) = \frac{1}{1 - \sigma_t}(\mathbf{x}_t - \mathbf{x}_0). \tag{3}$$

Group Relative Policy Optimization (GRPO). Given a prompt $\mathbf{c}$, GRPO samples a group of G rollouts and optimizes the diffusion policy π_θ using a group-normalized advantage:

$$\mathcal{L}_{GRPO}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} A_i \cdot \sum_{t=1}^{T} \log \pi_\theta(\mathbf{x}_{t-1}^{(i)} | \mathbf{x}_t^{(i)}, \mathbf{c}), \tag{4}$$

where T is the number of diffusion steps. We defer the reward composition, advantage normalization, and the stochastic rollout formulation to Appendix 5 and only keep the key equations in the corresponding modules.

3.2 SAGE-GRPO FRAMEWORK

3.2.1 MICRO-LEVEL EXPLORATION: PRECISE SDE AND GRADIENT EQUALIZATION

To enable stochastic exploration for GRPO, we perturb Rectified Flow with a marginal-preserving SDE whose noise stays aligned with the video manifold $\mathcal{M} \subset \mathbb{R}^D$ (Figure 2). The key challenge is computing the correct noise standard deviation $\Sigma_t^{1/2}$ during discrete SDE discretization. For a

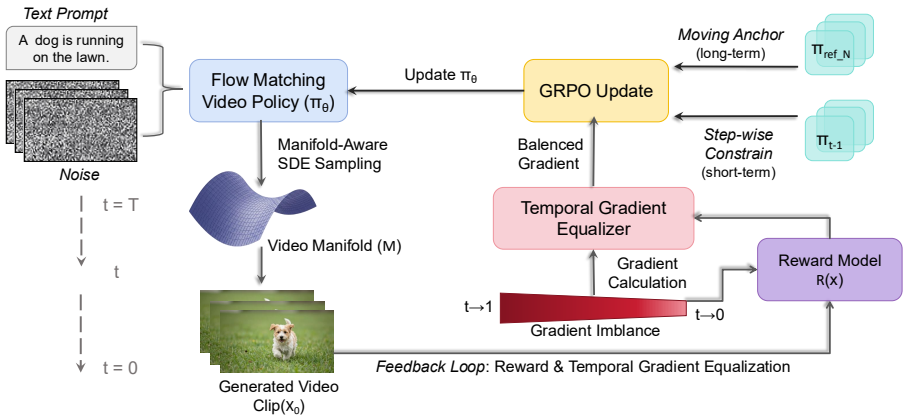


Figure 4: **The SAGE-GRPO Framework.** Our method resolves the stability-plasticity dilemma with three coupled components: **(Left)** a manifold-aware SDE that keeps exploration noise tangent to the video manifold, **(Middle)** a Temporal Gradient Equalizer that balances optimization across timesteps, and **(Right)** a Dual Trust Region that combines moving anchors and step-wise KL constraints for long-term stable alignment.

marginal-preserving SDE with diffusion coefficient $\varepsilon_t = \eta\sqrt{\sigma_t/(1-\sigma_t)}$, we integrate the variance over the interval $[\sigma_{t+1}, \sigma_t]$:

$$\Sigma_t = \int_{\sigma_{t+1}}^{\sigma_t} \varepsilon_s^2 \, \mathrm{d}s = \eta^2 \left[-(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right) \right], \tag{5}$$

where η is the exploration scaling factor. The logarithmic term accounts for the geometric contraction of the signal coefficient $(1-\sigma_t)$, which linear approximations fail to capture. Taking the square root yields the noise standard deviation:

$$\Sigma_t^{1/2} = \eta\sqrt{-(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right)}. \tag{6}$$

Applying Euler-Maruyama discretization with timestep $\Delta t = \sigma_t - \sigma_{t+1}$:

$$\boxed{\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \mathbf{v}_\theta(\mathbf{x}_t, t)\Delta t + \frac{\Sigma_t}{2}\mathbf{s}_\theta(\mathbf{x}_t) + \Sigma_t^{1/2}\boldsymbol{\epsilon},} \tag{7}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ injects stochasticity, $\mathbf{s}_\theta(\mathbf{x}_t) \approx -(\mathbf{x}_t - \hat{\mathbf{x}}_0)/\sigma_t^2$ is the score function estimate. Since Σ_t the integrated variance is over $[\sigma_{t+1}, \sigma_t]$, the stochastic term is used $\Sigma_t^{1/2}$ directly without an additional $\sqrt{\Delta t}$ factor. The Itô correction term $\frac{\Sigma_t}{2}\mathbf{s}_\theta(\mathbf{x}_t)$ ensures consistency with Rectified Flow marginals; a detailed derivation is provided in Appendix A.

As shown in Figure 2, our method creates a smaller, manifold-aligned exploration region (blue ellipsoid) that stays tangent to the flow trajectory, whereas conventional methods create larger, off-manifold exploration regions (red sphere) that cause state drift. This geometric insight ensures that every exploration step remains within the legal video space, preventing temporal artifacts. Even with correct noise injection, the diffusion process has an inherent signal-to-noise imbalance across timesteps: gradient norms vary by orders of magnitude (Figure 3), following a variance-gradient inverse relationship. For a Gaussian transition $\pi(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_t\mathbf{I})$:

$$\|\nabla_{\boldsymbol{\mu}} \log \pi\| \propto \frac{1}{\Sigma_t^{1/2}}, \tag{8}$$

causing gradients to vanish at high noise ($t \to 1$) and explode at low noise ($t \to 0$), biasing learning toward certain phases. To counteract this imbalance, we estimate a per-timestep gradient scale $\mathcal{N}_t$ from the SDE parameters (Appendix E) and apply a robust normalization:

$$\boxed{S_t = \frac{\text{Median}(\{\mathcal{N}_\tau\}_{\tau=1}^T)}{\mathcal{N}_t + \epsilon},} \tag{9}$$



Figure 5: **Qualitative comparison against baselines.** Three prompts illustrate our core gains: (Top) Reduced temporal jitter while preserving accurate visual contents; (Middle) Enhanced alignment and photorealism under occlusion and lighting changes; (Bottom) Stronger semantic alignment with consistent prompt matching across frames.

where ϵ is a small constant. This equalization normalizes optimization pressure across timesteps so that structural and textural updates contribute equally; details in Figure 6 and Appendix E.

GRPO With Composite Reward and Group-Normalized Advantage. We score each rollout $\mathbf{x}_0$ by a composite reward $R(\mathbf{x}_0)$ and compute the group-normalized advantage A_i:

$$A_i = \frac{r_i - \mu_R}{\sigma_R + \epsilon}, \tag{10}$$

where $r_i = R(\mathbf{x}_0^{(i)})$, $\mu_R = \frac{1}{G}\sum_{j=1}^{G} r_j$, and $\sigma_R^2 = \frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_R)^2$. Full definitions and implementation-aligned details are in Appendix D.

3.2.2 Macro-Level Exploration: Dual Trust Region Optimization

With micro-level exploration stabilized, we aim to prevent the policy model from drifting away from the data manifold and getting stuck in off-manifold local optima (Figure 2). We frame KL divergence as a dynamic anchoring mechanism that constrains exploration towards the data manifold.

KL Divergence as Dynamic Anchor. For a Gaussian policy $\pi(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_t \mathbf{I})$, the KL divergence between the current policy π_θ and a reference policy π_{ref} is:

$$D_{KL}(\pi_\theta||\pi_{ref}) = \mathbb{E}_{\mathbf{x}_t\sim\pi_\theta}\left[\frac{(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{ref})^2}{2\Sigma_t^2}\right] \approx \frac{(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{ref})^2}{2\Sigma_t^2}, \tag{11}$$

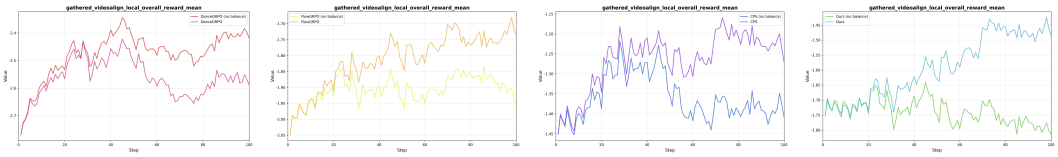
where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\mu}_{ref}$ are the mean predictions of the current and reference policies, respectively. KL divergence acts as a distance metric in policy space, anchoring the current policy to the reference. The choice of reference determines the constraint nature: a fixed reference creates a hard constraint, while a moving reference enables adaptive exploration.

Fixed KL: Hard Constraint Limiting Optimality. Traditional approaches anchor the policy to a fixed pretrained reference $\pi_{ref} = \pi_0$ using the constraint $D_{KL}(\pi_\theta||\pi_0)$. However, if the optimal policy π^* is far from π_0, this rigid penalty restricts long-term exploration and prevents convergence to π^*, resulting in underfitting.

Step-wise KL: Velocity Constraint. Step-wise KL uses the previous step's policy as reference: $\pi_{ref} = \pi_{k-1}$, where k denotes the optimization step. This constraint $D_{KL}(\pi_\theta||\pi_{k-1})$ acts as a velocity limit, restricting the magnitude of parameter updates per step:

$$\|\nabla_\theta D_{KL}(\pi_\theta||\pi_{k-1})\| \propto \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{k-1}\|/\Sigma_t, \tag{12}$$

ensuring smooth local transitions. However, velocity control alone only limits the magnitude of ∇_θ (the update direction) but does not bound the cumulative displacement $\|\theta_k - \theta_0\|$ from the initial



(a) DanceGRPO (b) FlowGRPO (c) CPS (d) Ours

Figure 6: **Temporal gradient balancing ablation across SDE formulations.** VideoAlign reward curves with and without the Gradient Norm Equalizer. Without balancing, low-noise timesteps dominate the optimization, leading to unstable or plateaued rewards. Equalization ensures smooth convergence and restricts gradient scale variance from over an order of magnitude to a constant factor.

parameters. This allows unbounded drift: the policy move slowly but consistently away from the manifold, eventually leading to degradation or reward hacking.

Periodical Moving KL: Position Control via Dynamic Trust Region. To counteract drift while maintaining plasticity, we introduce **Periodical Moving KL** that uses a periodically updated reference policy $\pi_{ref} = \pi_{k-N}$, where N is the update interval. For every N optimization step, we update the reference model: $\pi_{ref} \leftarrow \pi_\theta$, creating a resetting anchor mechanism. This allows the model to perform local exploration within N steps, then establish the new position as a safe region:

$$D_{KL}(\pi_\theta||\pi_{ref_N}) = \frac{(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{ref_N})^2}{2\Sigma_t^2}, \tag{13}$$

where $\boldsymbol{\mu}_{ref_N}$ is the reference mean prediction from N steps ago. This periodically resetting trust region acts as a multi-stage TRPO (Schulman et al., 2015), balancing step-wise reward improvements (plasticity) with distributional alignment (stability).

Dual KL: Position-Velocity Controller. We combine these two mechanisms into a **dual KL** objective that provides both position and velocity control:

$$\mathcal{L}_{KL} = \beta_{pos} \cdot D_{KL}(\pi_\theta||\pi_{ref_N}) + \beta_{vel} \cdot D_{KL}(\pi_\theta||\pi_{k-1}), \tag{14}$$

where β_{pos} and β_{vel} are weighting coefficients. The position term $D_{KL}(\pi_\theta||\pi_{ref_N})$ provides the primary directional anchor, preventing long-term drift by constraining the policy to remain within a reasonable distance from a recent valid distribution. The velocity term $D_{KL}(\pi_\theta||\pi_{k-1})$ acts as a damping factor, smoothing instantaneous updates and preventing abrupt policy changes. In practice, we compute the step-wise KL using log-probability differences from the rollout phase:

$$D_{KL}(\pi_\theta||\pi_{k-1}) \approx \mathbb{E}[\log \pi_{k-1}(\mathbf{x}_{t-1}|\mathbf{x}_t) - \log \pi_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)], \tag{15}$$

where the expectation is taken over samples generated with the previous policy π_{k-1}. The full SAGE-GRPO objective that combines the GRPO policy loss, temporal equalization, and Dual KL regularization is provided in Appendix F.

4 Experiments

4.1 Experimental Setup

Implementation Details. *Training.* We train all GRPO variants on HunyuanVideo 1.5 (Wu et al., 2025) with per-GPU batch size 2 and 4 gradient accumulation steps (effective batch size 8); each video contains 81 frames, and we apply a GRPO update every 20 sampling step along the diffusion trajectory. The fully on-policy training samples from ∼10k cinematic text prompts without requiring ground-truth videos (Appendix I). We evaluate SAGE-GRPO against DanceGRPO (Xue et al., 2025), FlowGRPO (Liu et al., 2025b), and CPS (Wang & Yu, 2025) using a unified infrastructure, scheduling the KL weight as $\lambda_{KL} \in [10^{-7}, 10^{-5}]$ (Appendix F). *Reward models.* Our primary oracle is VideoAlign (Liu et al., 2025c), combining Visual Quality (VQ), Motion Quality (MQ), and Text Alignment (TA) into a composite reward $R = w_{vq}S_{vq} + w_{mq}S_{mq} + w_{ta}S_{ta}$. To further explore generalization ability (Section 4.2), we conduct experiments on HunyuanVideo1.5, SD3.5-medium (Esser et al., 2024), and FLUX.1-dev (Labs et al., 2025) using the HPSv3 (Ma et al., 2025b) reward. *Note on PickScore reporting.* For readability, all reported PickScore values in this paper (main tables, appendix tables, and supplementary CSVs) are the raw PickScore multiplied by 26, so

Table 2: **Main Comparison on Video Generation Benchmarks.** Comparison of SAGE-GRPO with baselines under two reward settings. For each method, we report results *without* KL (**w/o KL**) and *with* fixed KL constraints (**w/ Fixed KL**). For SAGE-GRPO, we demonstrate the **w/ Dual Moving KL** mechanism. **Bold**, underline, and gray indicate the best, second-best, and third-best results.

Method	Configuration	VideoAlign Metrics				Visual Metrics	
		Overall	VQ	MQ	TA	CLIPScore	PickScore
HunyuanVideo 1.5 (Original)	-	0.0654	-0.7539	-0.5870	1.4063	0.5409	19.2322
Setting A: Averaged Rewards ($w_{vq} = 1.0, w_{mq} = 1.0, w_{ta} = 1.0$)							
DanceGRPO	w/o KL	0.2768	-0.7589	-0.3852	1.4209	0.5386	19.1828
	w/ Fixed KL	0.0979	-0.8077	-0.5091	1.4147	0.5403	19.1230
FlowGRPO	w/o KL	0.2733	-0.7151	-0.5286	1.5170	0.5443	19.2244
	w/ Fixed KL	0.1880	-0.6771	-0.5912	1.4563	0.5431	19.2582
CPS	w/o KL	0.6343	-0.4855	-0.4021	1.5219	0.5479	19.2712
	w/ Fixed KL	0.0928	-0.7156	-0.5825	1.3908	0.5479	19.1594
SAGE-GRPO	w/o KL	0.4859	-0.6104	-0.4141	1.5104	0.5423	19.1360
	w/ Fixed KL	0.2244	-0.7438	-0.5320	1.5001	0.5446	19.1932
	w/ Dual Mov KL	0.2173	-0.7881	-0.4249	1.4303	0.5430	**19.3752**
Setting B: Alignment-Focused ($w_{vq} = 0.5, w_{mq} = 0.5, w_{ta} = 1.0$)							
DanceGRPO	w/o KL	-0.2172	-0.8854	-0.6218	1.2901	0.5439	19.1152
	w/ Fixed KL	0.1290	-0.7739	-0.5083	1.4112	0.5452	18.9176
FlowGRPO	w/o KL	0.4773	-0.5671	-0.4731	1.5175	0.5403	19.1074
	w/ Fixed KL	0.2103	-0.6654	-0.5506	1.4263	0.5427	19.2608
CPS	w/o KL	0.3694	-0.6650	-0.5325	**1.5669**	0.5479	19.0086
	w/ Fixed KL	0.3705	-0.6121	-0.4787	1.4613	0.5458	19.1464
SAGE-GRPO	w/o KL	-0.1222	-0.8720	-0.6046	1.3544	0.5404	19.1282
	w/ Fixed KL	0.2857	-0.7062	-0.4425	1.4344	0.5414	19.1802
	w/ Dual Mov KL	**0.8066**	**-0.4765**	**-0.2384**	1.5216	**0.5484**	19.2920

that score differences between methods are visible at the second decimal rather than buried in the fourth; underlying raw scores can be recovered by dividing by 26. *Validation.* We use 32 held-out prompts (*realistic* n_p=14, *animated* n_p=10, *abstract* n_p=8) for qualitative analysis and user studies (Section 4.3 4.4). For image generalization, we evaluate SD3.5 and FLUX at three consecutive checkpoints around 1500 training steps (steps $\{1440, 1500, 1560\}$ for both backbones) over 256 prompts each, giving $3 \times 256 = 768$ prompt evaluations per cell.

4.2 Main Results

We use two reward configurations (Table 2): **averaged** (w_{vq}=w_{mq}=w_{ta}=1.0) and **alignment-focused** (w_{vq}=w_{mq}=0.5, w_{ta}=1.0), both with frozen VideoAlign as evaluator and a unified training infrastructure that varies only the GRPO algorithm. Under the averaged setting (matching Longcat-Video (Team et al., 2025)) adding KL typically improves visual quality but slightly degrades Overall, consistent with reward hacking (Li et al., 2025b); the alignment-focused setting provides a more reliable optimization target. SAGE-GRPO + Dual Moving KL achieves the best Overall, VQ, MQ, and CLIPScore on Setting B while remaining close to the best TA, and is the only method that improves Overall over the HunyuanVideo1.5 baseline in both settings. A direct head-to-head against the concurrent TempFlow-GRPO (He et al., 2025b) is in Appendix J.

Generalization across reward models and backbones (HPSv3). We re-train all four GRPO variants under the HPSv3 (Ma et al., 2025b) reward on HunyuanVideo1.5 (video), SD3.5-medium (Esser et al., 2024), and FLUX.1-dev (Labs et al., 2025) (training and validation protocols in §4.1). Per-prompt paired t-tests backing the SD3.5 / FLUX rows are in Appendix K.

4.3 Ablation and Qualitative Analysis

Ablations: SDE, equalizer, KL, moving frequency N. The four *w/ Fixed KL* rows in Table 2 share identical training infrastructure, KL regularization, and reward configuration, and differ *only* in the SDE noise-injection formula (Table 1); these four rows therefore directly serve as the per-method SDE ablation. Under Setting B, our Precise Manifold-Aware SDE improves Overall reward over DanceGRPO ($0.1290 \rightarrow 0.2857$), FlowGRPO ($0.2103 \rightarrow 0.2857$), and CPS ($0.3705 \rightarrow \mathbf{0.8066}$ once further coupled with Dual Moving KL). Figure 6 reports the gradient-equalizer effect across SDE formulations, and KL-strategy / KL-weight ablations (Appendix Figures 9–10, qualitative in

Table 3: **HPSv3-trained generalization across video and image backbones.** Methods are retrained under HPSv3 (Ma et al., 2025b) reward and evaluated on the training reward (HPSv3) plus two visual metrics (PickScore, CLIPScore). Section bands distinguish the *video* backbone (HunyuanVideo1.5) from the two *image* backbones (SD3.5-medium, FLUX.1-dev); image rows report mean ± SEM. **Bold**, underline, and gray indicate the best, second best, and third best within each backbone band.

Method	Training Reward HPSv3 ↑	Visual Metrics PickScore ↑	Visual Metrics CLIPScore ↑
Video Backbone: HunyuanVideo 1.5			
DanceGRPO	4.471	**18.876**	0.584
FlowGRPO	4.739	18.798	0.626
CPS	4.736	18.642	0.626
SAGE-GRPO	**5.500**	18.824	**0.627**
Image Backbone: SD3.5-medium			
DanceGRPO	9.237 ± 0.112	21.762 ± 0.052	**0.970 ± 0.005**
FlowGRPO	8.416 ± 0.116	21.606 ± 0.052	0.958 ± 0.006
CPS	**9.438 ± 0.110**	**21.840 ± 0.052**	0.966 ± 0.005
SAGE-GRPO	9.376 ± 0.114	**21.840 ± 0.052**	**0.970 ± 0.005**
Image Backbone: FLUX.1-dev			
DanceGRPO	**10.20 ± 0.099**	**22.126 ± 0.052**	**0.917 ± 0.006**
FlowGRPO	10.07 ± 0.099	22.048 ± 0.052	0.914 ± 0.006
CPS	10.14 ± 0.099	22.100 ± 0.052	0.914 ± 0.006
SAGE-GRPO	10.17 ± 0.099	22.100 ± 0.052	**0.917 ± 0.006**

Appendix Figures 11–12; Appendix F and Appendix H) confirm the best schedule and moving frequency N of Dual Moving KL.

Qualitative analysis. Figure 5 shows gains in coherence, photorealism, and semantic alignment over baselines (extended in Appendix Figures 8, 11–12), details in Section G.

4.4 User Study

To validate the quality aligning with human preference, we conducted a user study involving 29 evaluators across 32 prompts. We compared SAGE-GRPO against the baselines (all at iteration 100, step 40, Setting B) in terms of Visual Quality, Motion Quality, and Semantic Alignment. Table 4 reports the resulting pairwise win rates of SAGE-GRPO against each baseline.

Table 4: **User Preference Study: aggregated win rates** ($29\times32\times3 = 2784$ comparisons). All entries are pairwise win rates of SAGE-GRPO against the named baseline.

SAGE-GRPO vs.	Visual Quality	Motion Quality	Semantic Alignment
DanceGRPO	85.9%	75.8%	79.2%
FlowGRPO	83.8%	79.2%	71.9%
CPS	80.2%	70.8%	67.9%

5 Conclusion

We presented **SAGE-GRPO**, a manifold-aware GRPO framework for stable alignment of video generation. At the micro-level, a Precise Manifold-Aware SDE keeps exploration near the manifold trajectory, and a Gradient Norm Equalizer balances optimization updates across timesteps. At the macro-level, a Dual Trust Region uses position and velocity controls to prevent off-manifold drift and maintain sustained plasticity. Experiments on HunyuanVideo1.5 with VideoAlign, alongside generalization tests on SD3.5-medium and FLUX.1-dev using HPSv3 demonstrates consistent improvements in reward maximization and overall visual quality compared to existing baselines.

AI USE STATEMENT

We use Google Gemini 2.5 Pro to generate the $\sim$10k cinematic text prompts that drive the on-policy GRPO rollouts; this LLM-generated prompt pool directly affects the data distribution our policy is optimized over and is therefore declared in the paper. The generation template, intent (rich physical detail to give the reward model enough surface to score policy changes), and release plan are described in Appendix I. The core SAGE-GRPO method itself (Precise Manifold-Aware SDE, Gradient Norm Equalizer, Dual Trust Region) is an SDE-level and KL-level reformulation of GRPO and does not use LLMs as an algorithmic component. Generative AI was also used to adapt the LaTeX formatting to the ICLR 2027 template.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2403.03206.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Alexander Gambashidze, Anton Kulikov, Yuriy Sosnin, and Ilya Makarov. Aligning diffusion models with noise-conditioned perception. *arXiv preprint arXiv:2406.17636*, 2024.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dailan He, Guanlin Feng, Xingtong Ge, Yazhe Niu, Yi Zhang, Bingqi Ma, Guanglu Song, Yu Liu, and Hongsheng Li. Neighbor grpo: Contrastive ode policy optimization aligns flow models. *arXiv preprint arXiv:2511.16955*, 2025a.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv: 2508.04324*, 2025b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tao Huang, Guangqi Jiang, Yanjie Ze, and Huazhe Xu. Diffusion reward: Learning rewards via conditional video diffusion. In *European Conference on Computer Vision*, pp. 478–495. Springer, 2024.

Zhiwei Jia, Yuesong Nan, Huixi Zhao, and Gengdai Liu. Reward fine-tuning two-step diffusion models via learning differentiable latent-space surrogate reward. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12912–12922, 2025.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Zhen Li, Bo Zhang, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

Dongyang Jin, Ryan Xu, Jianhao Zeng, Rui Lan, Yancheng Bai, Lei Sun, and Xiangxiang Chu. Semantic context matters: Improving conditioning for autoregressive models. *arXiv preprint arXiv:2511.14063*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv: 2506.15742*, 2025.

Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Dongyang Jin, Ryan Xu, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*, 2025.

Jiachen Li, Weixi Feng, Wenhu Chen, and William Yang Wang. Reward guided latent consistency distillation. *arXiv preprint arXiv:2403.11027*, 2024.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Rui Li, Yuanzhi Liang, Ziqi Ni, Haibing Huang, Chi Zhang, and Xuelong Li. Growing with the generator: Self-paced grpo for video generation. *arXiv preprint arXiv:2511.19356*, 2025b.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025c.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. *arXiv preprint arXiv:2512.21514*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025b.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025c.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Zeqian Long, Mingzhe Zheng, Kunyu Feng, Xinhua Zhang, Hongyu Liu, Harry Yang, Linfeng Zhang, Qifeng Chen, and Yue Ma. Follow-your-shape: Shape-aware image editing via trajectory-guided region control. *arXiv preprint arXiv:2508.08134*, 2025.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025a.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, October 2025b.

Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, et al. Video generation models are good latent reward models. *arXiv preprint arXiv:2511.21541*, 2025.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiangwei Shen, Zhimin Li, Zhantao Yang, Shiyi Zhang, Yingfang Zhang, Donghao Li, Chunyu Wang, Qinglin Lu, and Yansong Tang. Directly aligning the full diffusion trajectory with fine-grained human preference. *arXiv preprint arXiv:2509.06942*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2020a.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, S. Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference On Learning Representations*, 2020b.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025.

Fu-Yun Wang, Ling Yang, Zhaoyang Huang, Mengdi Wang, and Hongsheng Li. Rectified diffusion: Straightness is not your need in rectified flow. *arXiv preprint arXiv:2410.07303*, 2024.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

Ryan Xu, Dongyang Jin, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, and Xiangxiang Chu. Scalar: Scale-wise controllable visual autoregressive learning. *arXiv preprint arXiv:2507.19946*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Xudong Yu, Chenjia Bai, Haoran He, Changhong Wang, and Xuelong Li. Regularized conditional diffusion model for multi-task preference alignment. *Advances in Neural Information Processing Systems*, 37:139968–139996, 2024.

Shengjun Zhang, Zhang Zhang, Chensheng Dai, and Yueqi Duan. E-grpo: High entropy steps drive effective reinforcement learning for flow models. *arXiv preprint arXiv:2601.00423*, 2026.

Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. *arXiv preprint arXiv:2502.01051*, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv: 2509.16117*, 2025.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024.

Yujie Zhou, Pengyang Ling, Jiazi Bu, Yibin Wang, Yuhang Zang, Jiaqi Wang, Li Niu, and Guangtao Zhai. Fine-grained grpo for precise preference alignment in flow models. *arXiv preprint arXiv:2510.01982*, 2025.

BROADER IMPACT STATEMENT

This paper presents a method for more stable reinforcement learning alignment of text-to-video generation models. By improving temporal consistency and text alignment under a fixed reward model, our work may strengthen creative tools, scientific communication, and educational content that rely on controllable video synthesis. At the same time, stronger video generation systems can exacerbate existing concerns about misinformation, deepfakes, biased or harmful content, and the computational cost of large-scale training and sampling. Our experiments are conducted in a research setting on an existing model and evaluator, and our user study involves 29 voluntary evaluators rating 32 prompts comparing SAGE-GRPO against baselines in terms of visual quality, motion quality, and semantic alignment; there is no collection of personal data, but any future deployment should include safeguards such as content moderation, dataset auditing, and human oversight to reduce these risks.

A DERIVATION OF MANIFOLD-AWARE SDE VARIANCE

Derivation To enable stochastic exploration for GRPO, we need to convert the deterministic Rectified Flow ODE into a stochastic differential equation (SDE) that preserves the marginal probability distribution at each timestep. Recall the general form of a marginal-preserving SDE for flow matching:

$$\mathrm{d}\mathbf{z}_t = \left(\mathbf{v}_\theta(\mathbf{x}_t, t) - \tfrac{1}{2}\varepsilon_t^2 \mathbf{s}_\theta(\mathbf{x}_t)\right)\mathrm{d}t + \varepsilon_t\,\mathrm{d}\mathbf{w}_t, \tag{16}$$

where ε_t is the diffusion coefficient (a function of t), $\mathbf{w}_t$ is a Brownian motion, and $\mathbf{s}_\theta(\mathbf{x}_t) \approx -(\mathbf{x}_t - \hat{\mathbf{x}}_0)/\sigma_t^2$ is the score function estimate. The Itô correction term $\frac{1}{2}\varepsilon_t^2\mathbf{s}_\theta(\mathbf{x}_t)$ ensures that the SDE preserves the same marginal distribution as the deterministic ODE.

To discretize this SDE, we assume that $\mathbf{v}_\theta(\mathbf{x}_t, t)$ and $\mathbf{s}_\theta(\mathbf{x}_t)$ remain approximately constant during the interval $[\sigma_{t+1}, \sigma_t]$, where σ_t is the noise level at timestep t. The key challenge is to compute the integrated variance Σ_t for the stochastic term. We define:

$$\Sigma_t \coloneqq \int_{\sigma_{t+1}}^{\sigma_t} \varepsilon_s^2\,\mathrm{d}s. \tag{17}$$

For Rectified Flow, we choose $\varepsilon_t = \eta\sqrt{\frac{\sigma_t}{1-\sigma_t}}$ to match the geometric structure of the flow trajectory, where η is the exploration scaling factor. Substituting this form and integrating:

$$\Sigma_t = \int_{\sigma_{t+1}}^{\sigma_t} \eta^2 \frac{\sigma_s}{1-\sigma_s}\,\mathrm{d}s \tag{18}$$

$$= \eta^2 \int_{\sigma_{t+1}}^{\sigma_t} \left(\frac{1}{1-\sigma_s} - 1\right)\mathrm{d}s \tag{19}$$

$$= \eta^2 \left[-\log(1-\sigma_s) - \sigma_s\right]_{\sigma_{t+1}}^{\sigma_t} \tag{20}$$

$$= \eta^2 \left[-(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right)\right]. \tag{21}$$

Taking the square root, we obtain the noise standard deviation:

$$\Sigma_t^{1/2} = \eta\sqrt{-(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right)}. \tag{22}$$

The logarithmic term $\log((1-\sigma_{t+1})/(1-\sigma_t))$ accounts for the geometric contraction of the signal coefficient $(1-\sigma_t)$, which linear approximations fail to capture.

Applying Euler-Maruyama discretization with timestep $\Delta t = \sigma_t - \sigma_{t+1}$, the discretized SDE becomes:

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \mathbf{v}_\theta(\mathbf{x}_t, t)\Delta t + \frac{\Sigma_t}{2}\mathbf{s}_\theta(\mathbf{x}_t) + \Sigma_t^{1/2}\boldsymbol{\epsilon}, \tag{23}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the injected stochasticity. Note that Σ_t is already the integrated variance over the interval $[\sigma_{t+1}, \sigma_t]$, so the stochastic term uses $\Sigma_t^{1/2}$ directly without an additional $\sqrt{\Delta t}$ factor.

Problem Formulation. Let the noise level at timestep t be σ_t. In a Rectified Flow setting, the trajectory connects pure noise ($\sigma = 1$) to data ($\sigma = 0$). We aim to find the precise variance Σ_t required for the stochastic step such that the marginal distribution is preserved up to the second order.

Throughout this derivation, $\log$ denotes the natural logarithm. Let $\Delta\sigma = \sigma_t - \sigma_{t+1} > 0$. We analyze the terms inside the square root of our proposed Eq. 7. Let V_t denote the variance term:

$$V_t = -(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right) \tag{24}$$

Non-negativity of the integrated variance. For a non-increasing noise schedule satisfying $0 \le \sigma_{t+1} \le \sigma_t < 1$, define

$$V_t = -(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right), \qquad \Sigma_t = \eta^2 V_t. \tag{25}$$

Although the first term in V_t is non-positive, the logarithmic term always compensates for it. Specifically,

$$\log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right) = \int_{\sigma_{t+1}}^{\sigma_t} \frac{1}{1-u}\,\mathrm{d}u \tag{26}$$

$$\ge \int_{\sigma_{t+1}}^{\sigma_t} 1\,\mathrm{d}u = \sigma_t - \sigma_{t+1}, \tag{27}$$

where the inequality follows from $(1-u)^{-1} \ge 1$ for $u \in [0, 1)$. Equivalently,

$$\boxed{V_t = \int_{\sigma_{t+1}}^{\sigma_t} \frac{u}{1-u}\,\mathrm{d}u \ge 0.} \tag{28}$$

The inequality is strict whenever $\sigma_{t+1} < \sigma_t$, while $V_t = 0$ when the two endpoints coincide. Therefore, for $\eta \ge 0$, the noise standard deviation

$$\Sigma_t^{1/2} = \eta\sqrt{-(\sigma_t - \sigma_{t+1}) + \log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right)} \tag{29}$$

is real-valued throughout the stated domain. This result holds for any admissible interval and does not rely on a Taylor approximation or a small-step assumption. The endpoint $\sigma_t = 1$ is excluded because the unregularized variance integral diverges for any $\sigma_{t+1} < 1$; numerical boundary regularization must therefore be treated separately.

Taylor Expansion Analysis. First, we express the logarithmic term using $\Delta\sigma$:

$$\log\left(\frac{1-\sigma_{t+1}}{1-\sigma_t}\right) = \log\left(\frac{1-(\sigma_t - \Delta\sigma)}{1-\sigma_t}\right) = \log\left(1 + \frac{\Delta\sigma}{1-\sigma_t}\right) \tag{30}$$

Let $x = \frac{\Delta\sigma}{1-\sigma_t}$. Let $\Delta\sigma = \sigma_t - \sigma_{t+1}$ and $x = \Delta\sigma/(1-\sigma_t)$. For $x \ll 1$, the logarithmic term admits the second-order expansion $\log(1+x) = x - \frac{1}{2}x^2 + \mathcal{O}(x^3)$:

$$\log\left(1 + \frac{\Delta\sigma}{1-\sigma_t}\right) \approx \frac{\Delta\sigma}{1-\sigma_t} - \frac{1}{2}\left(\frac{\Delta\sigma}{1-\sigma_t}\right)^2 \tag{31}$$

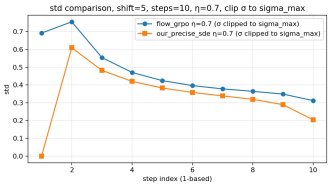
Substituting this back into V_t:

$$V_t \approx -\Delta\sigma + \left(\frac{\Delta\sigma}{1-\sigma_t} - \frac{1}{2}\frac{\Delta\sigma^2}{(1-\sigma_t)^2}\right) \tag{32}$$

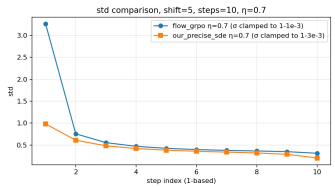
$$= \Delta\sigma\left(\frac{1}{1-\sigma_t} - 1\right) - \frac{1}{2}\frac{\Delta\sigma^2}{(1-\sigma_t)^2} \tag{33}$$

$$= \Delta\sigma\left(\frac{\sigma_t}{1-\sigma_t}\right) - \mathcal{O}(\Delta\sigma^2) \tag{34}$$

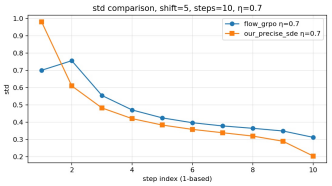
The leading term $\Delta\sigma\frac{\sigma_t}{1-\sigma_t}$ represents the ideal variance scaling for a geometric schedule, which linear approximations fail to capture.



(a) Both using FlowGRPO's σ schedule.



(b) Both clamped at $1 - 3 \times 10^{-3}$.



(c) Each using default implementation.

Figure 7: **Step-wise std comparison: our precise SDE vs. FlowGRPO.** (a) When both use FlowGRPO's σ schedule, our integral-based formulation yields near-zero std at the first step due to equal endpoints. (b) When both are clamped at $(1 - \sigma) \geq 3 \times 10^{-3}$, FlowGRPO explodes at step 1, while ours remains stable. (c) Under default implementations, ours maintains lower std across most steps. This supports that we remove ineffective high-frequency noise and explore more precisely along the manifold (Section 3.2.1).

B Standard Deviation Comparison: Ours vs. FlowGRPO

Figure 7 compares the noise standard deviation per step between our precise SDE and FlowGRPO under three parameterization regimes to understand how different noise handling strategies affect exploration behavior.

Regime (a): Both methods using FlowGRPO's σ schedule. When both methods use FlowGRPO's default σ schedule (where σ_t is set to $\sigma_{\max}$ at early steps), our precise SDE exhibits near-zero standard deviation at the first step. This occurs because our method computes noise variance via integration: $\Sigma_t = \int_{\sigma_{t+1}}^{\sigma_t} \varepsilon_s^2 \, \mathrm{d}s$. When both endpoints are equal ($\sigma_t = \sigma_{t+1} = \sigma_{\max}$), the integration interval collapses, and the logarithmic term $\log((1 - \sigma_{t+1})/(1 - \sigma_t))$ evaluates to zero, yielding $\Sigma_t \approx 0$. This demonstrates that our integral-based formulation is sensitive to the σ schedule and requires proper boundary handling.

Regime (b): Both methods using aggressive clamping at $1 - 3 \times 10^{-3}$. When both methods apply the same clamping threshold $(1 - \sigma) \geq 3 \times 10^{-3}$ (equivalently, $\sigma \leq 1 - 3 \times 10^{-3}$), FlowGRPO exhibits explosive behavior at the first step, with standard deviation reaching values around 3.0. This instability arises because FlowGRPO's noise computation involves a ratio $\sigma/(1 - \sigma)$; when $(1 - \sigma)$ is clamped to a small constant while σ remains large, the denominator becomes artificially small, causing the ratio to explode. In contrast, our precise SDE maintains stable and controlled standard deviation throughout, starting around 1.0 and decaying smoothly, demonstrating that our manifold-aware formulation inherently handles low-noise regimes more robustly.

Regime (c): Each method using its default implementation. Under their respective default configurations, FlowGRPO uses its standard σ schedule, while our method applies clamping at $(1 - \sigma) \geq 3 \times 10^{-3}$. Our method maintains a lower standard deviation than FlowGRPO across most of the diffusion trajectory, particularly in later steps. This demonstrates that our precise SDE effectively reduces injected noise magnitude, leading to more refined exploration along the data manifold.

Across all three regimes, our method consistently achieves smaller or more stable standard deviation than FlowGRPO. This supports the main-figure narrative (Figure 1): we remove unnecessary high-frequency noise energy in high-noise regions, enabling more precise exploration that stays closer to the data manifold. This behavior aligns with the micro-level exploration design of our SDE in Section 3.2.1.

C Theoretical Gradient Norm Analysis

Here we derive the relationship between the gradient norm and the noise schedule. For a Gaussian policy $\pi(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta, \Sigma_t \mathbf{I})$, the gradient of the log-probability with respect to the drift parameter μ_θ is:

$$\nabla_\mu \log \pi = \frac{\mathbf{x}_{sample} - \mu_\theta}{\Sigma_t} \tag{35}$$



Prompt: ...a teenage boy sitting alone in a brightly lit coffee shop,...After savoring the taste, he gently lowers the mug back onto its saucer without a sound. He turns his head slightly, ...with a calm, contemplative expression.



Prompt: ...An older, white-haired chef in a professional kitchen. His face, etched with lines of fatigue, is illuminated by a single source of side lighting, casting deep shadows that accentuate his somber mood....

Figure 8: **Qualitative comparison highlighting emotional alignment.** Two prompts illustrate SAGE-GRPO's ability to better align with emotional descriptions: (Top) A teenage boy in a coffee shop, where SAGE-GRPO captures the "calm, contemplative expression" and gentle motion of lowering the mug, while baselines show neutral expressions and abrupt movements. (Bottom) An older chef in a kitchen, where SAGE-GRPO consistently renders the "lines of fatigue" and "somber mood" through deep side-lighting shadows, while baselines fail to convey the intended emotional depth. Our manifold-aware exploration enables precise alignment with subtle emotional and action cues.

Since $\mathbf{x}_{sample} \sim \mathcal{N}(\mu_\theta, \Sigma_t)$, the expected norm is proportional to the standard deviation of the noise:

$$\mathbb{E}[\|\nabla_\mu \log \pi\|] \propto \frac{\sqrt{\Sigma_t}}{\Sigma_t} = \frac{1}{\sqrt{\Sigma_t}} \tag{36}$$

Given our derived Manifold-Aware variance $\Sigma_t \approx \eta^2 \Delta\sigma \frac{\sigma_t}{1-\sigma_t}$, the gradient norm scales as:

$$\|\nabla\| \propto \sqrt{\frac{1-\sigma_t}{\sigma_t \Delta\sigma}} \tag{37}$$

This confirms that as $\sigma_t \to 0$ (low noise), the gradient norm explodes, necessitating our proposed Gradient Equalizer.

D GRPO Reward and Advantage Details

Here we provide the implementation-aligned definitions of reward composition and the group-normalized advantage used in Equation (4). Following VideoAlign (Liu et al., 2025c), we construct a composite reward for a generated video $\mathbf{x}_0$:

$$R(\mathbf{x}_0) = w_{vq} S_{vq}(\mathbf{x}_0) + w_{mq} S_{mq}(\mathbf{x}_0) + w_{ta} S_{ta}(\mathbf{x}_0), \tag{38}$$

where S_{vq}, S_{mq}, and S_{ta} score visual quality, motion quality, and text alignment, and w_{vq}, w_{mq}, w_{ta} are fixed scalar weights.

Given a prompt $\mathbf{c}$, GRPO samples a group of G rollouts $\{\mathbf{x}_0^{(i)}\}_{i=1}^G$ and computes rewards $r_i = R(\mathbf{x}_0^{(i)})$. We use the group mean and standard deviation as a baseline:

$$\mu_R = \frac{1}{G}\sum_{j=1}^{G} r_j, \qquad \sigma_R = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_R)^2}, \tag{39}$$

and define the normalized advantage:

$$A_i = \frac{r_i - \mu_R}{\sigma_R + \epsilon}, \tag{40}$$

where ϵ is a small constant for numerical stability.

E Temporal Gradient Equalizer: Derivation of $\mathcal{N}_t$

We outline how to obtain a per-timestep gradient scale proxy $\mathcal{N}_t$ that is compatible with the SDE transition used in Section 3.2.1. Consider a Gaussian transition $\pi(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_t \mathbf{I})$ parameterized through the network output (e.g., velocity/denoiser prediction) and a noise variance Σ_t determined by the chosen SDE. The log-probability gradient with respect to the mean parameter satisfies:

$$\nabla_{\boldsymbol{\mu}} \log \pi = \frac{\mathbf{x}_{sample} - \boldsymbol{\mu}_\theta}{\Sigma_t}. \tag{41}$$

Since $\mathbf{x}_{sample} - \boldsymbol{\mu}_\theta \sim \mathcal{N}(\mathbf{0}, \Sigma_t \mathbf{I})$, its magnitude is $\mathcal{O}(\Sigma_t^{1/2})$ in expectation, yielding the inverse relationship:

$$\mathbb{E}\big[\|\nabla_{\boldsymbol{\mu}} \log \pi\|\big] \propto \frac{1}{\Sigma_t^{1/2}}. \tag{42}$$

In practice, the network does not directly parameterize $\boldsymbol{\mu}_\theta$; instead, $\boldsymbol{\mu}_\theta$ is obtained by composing the network prediction with the SDE/solver update rule, introducing an additional sensitivity factor. Let λ_t denote the scalar sensitivity from the solver mapping (details depend on the SDE type and discretization). We use the proxy:

$$\mathcal{N}_t = \frac{\lambda_t}{\Sigma_t^{1/2}}, \tag{43}$$

and define the Temporal Gradient Equalizer (Equation (9)) as a robust normalization:

$$S_t = \frac{\mathrm{Median}(\{\mathcal{N}_\tau\}_{\tau=1}^T)}{\mathcal{N}_t + \epsilon}. \tag{44}$$

This produces approximately uniform gradient scales across timesteps, aligning with the empirical observation in Figure 3 and the training-curve improvement in Figure 6.

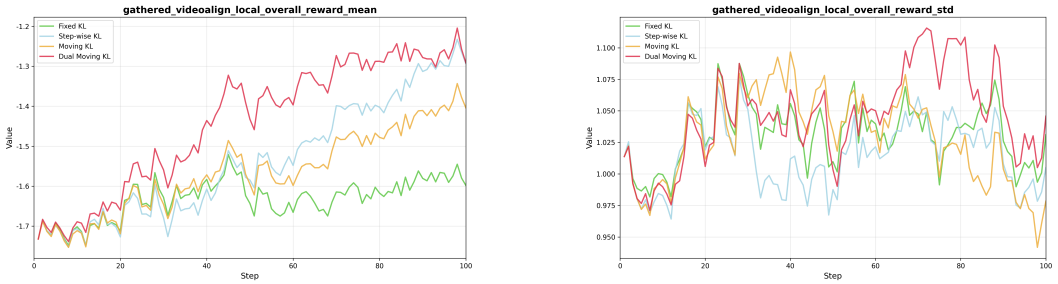
F SAGE-GRPO Objective and Adaptive KL Weighting

We provide the complete objective used in SAGE-GRPO, combining GRPO, the Temporal Gradient Equalizer, and Dual KL regularization, together with a principled schedule for the overall KL coefficient. At each optimization step, we sample a group of G rollouts and compute advantages $\{A_i\}_{i=1}^G$ as in Appendix D.

Dual KL regularizer. We use two reference policies to implement a position–velocity controller in policy space (Section 3.2.2). The regularizer is

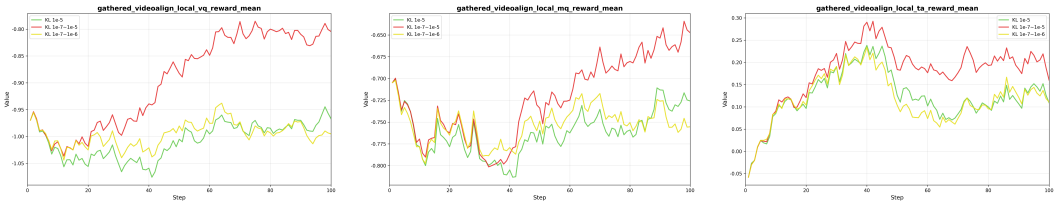
$$\mathcal{L}_{KL} = \beta_{pos} \cdot D_{KL}(\pi_\theta \| \pi_{ref_N}) + \beta_{vel} \cdot D_{KL}(\pi_\theta \| \pi_{k-1}), \tag{45}$$

where π_{k-1} is the previous policy and π_{ref_N} is a periodically refreshed anchor. The term $D_{KL}(\pi_\theta \| \pi_{k-1})$ constrains the *instantaneous update* (velocity control), while $D_{KL}(\pi_\theta \| \pi_{ref_N})$ constrains the *cumulative displacement* from the anchor (position control). This separation is important for long-horizon training: velocity-only constraints can still accumulate drift, whereas a single fixed anchor can be overly restrictive.



(a) Mean reward (b) Std (exploration)

Figure 9: **KL strategy ablation.** (a) Dual Moving KL achieves the highest and most stable reward, supporting the position-velocity control interpretation (Equation (14)). (b) Moving KL attains high exploration in early training steps but the exploration level drops in later stages. Dual Moving KL maintains a higher and more stable exploration level throughout training.



(a) VQ reward (b) MQ reward (c) TA reward

Figure 10: **KL weight ablation on VideoAlign rewards.** Comparison of three KL weight schedules: fixed 10^{-5} (green), two-stage $10^{-7} \to 10^{-5}$ (red), and two-stage $10^{-7} \to 10^{-6}$ (yellow). The two-stage schedule $10^{-7} \to 10^{-5}$ achieves the strongest and most consistent gains across VQ, MQ, and TA, supporting gradually increasing λ_{KL} to tighten the trust region (Appendix F).

Adaptive KL weighting. We interpret the overall KL coefficient λ_{KL} as a Lagrange multiplier associated with a trust-region constraint $\mathbb{E}[D_{KL}(\pi_\theta \| \pi_{ref})] \le \delta$. Instead of fixing λ_{KL}, we adapt it online so that the realized KL remains close to a target scale, analogous in spirit to adaptive behavior regularization in AWAC (Nair et al., 2020), where a temperature parameter is adapted from advantage statistics.

Warm-up (two-stage increase). Let $\lambda_{\min} = 10^{-7}$ and $\lambda_{\max} = 10^{-5}$ denote the minimum and maximum KL coefficients. During the first $K = 100$ optimization steps, we use a linear warm-up:

$$\lambda_{KL}(k) = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cdot \frac{k}{K}, \qquad k \le K, \tag{46}$$

which corresponds to the two-stage schedules reported in Figure 10. This design keeps the trust region weak early to avoid underfitting, and gradually strengthens it as the policy improves.

Conservative feedback control. After warm-up, we apply a proportional feedback update based on the recent KL history, similar to the P-term of a PID controller. Let $\bar{D}_{KL}$ be the mean of the last $H = 10$ observed KL values and let D_{target} be the desired KL scale. We define the KL error $e_{KL} = \bar{D}_{KL} - D_{target}$ and update

$$\lambda_{KL} \leftarrow \begin{cases} 0.9\,\lambda_{KL}, & \bar{D}_{KL} > (1 + 0.5)\, D_{target}, \\ 1.1\,\lambda_{KL}, & \bar{D}_{KL} < (1 - 0.5)\, D_{target}, \\ \lambda_{KL}, & \text{otherwise}, \end{cases} \qquad \lambda_{KL} \in [\lambda_{\min}, \lambda_{\max}], \tag{47}$$

where clipping enforces the same bounds as the warm-up stage. Intuitively, if the empirical KL is much larger than D_{target}, the controller reduces λ_{KL} to relax the constraint; if it is much smaller, the controller increases λ_{KL} to tighten the trust region. This combination of warm-up and feedback control stabilizes the effective trust-region radius and explains the smooth reward trajectories observed in Figure 10.



Prompt: …The soldier's features are etched with fatigue and grime; their skin pale against the dark dirt and sweat, their eyes hollow and bloodshot, their chins darkened by the rough stubble of days past…take a slow, measured sip, their throats noticeably moving as they swallow, and a look of desperate relief crosses their faces before they close their eyes.

Figure 11: **KL strategy ablation: qualitative comparison (Case 1).** Visual comparison across different KL strategies (no KL, Fixed KL, Step-wise KL, Moving KL, Dual Moving KL) on a prompt describing a fatigued soldier. Dual Moving KL produces more realistic facial details, better dirt and grime rendering, and maintains temporal consistency across frames compared to other variants.

Full SAGE-GRPO loss. Combining GRPO, the Temporal Gradient Equalizer, and the adaptively weighted Dual KL regularizer yields:

$$\boxed{\mathcal{L}_{SAGE\text{-}GRPO}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} A_i \cdot \sum_{t=1}^{T} S_t \cdot \log \pi_\theta(\mathbf{x}_{t-1}^{(i)} \mid \mathbf{x}_t^{(i)}, \mathbf{c}) - \lambda_{KL} \cdot \mathcal{L}_{KL}.} \tag{48}$$

G Additional Qualitative Results

Compute resources. All reported experiments were conducted on NVIDIA GPUs with at least 80 GB of memory, and all training runs used bf16 mixed-precision under fair comparison. The main experiments on HunyuanVideo 1.5 (Sections 4.2 and 4.2) were trained in a multi-node configuration with the per-GPU batch size, gradient accumulation, and rollout schedule reported in Section 4. The image generalization runs on SD3.5-medium and FLUX.1-dev (Section 4.2) were trained on a single H800 GPU per run, using the same training-loop interface as FlowGRPO (Liu et al., 2025b) for a strictly comparable infrastructure footprint. Inference and evaluation (VideoAlign, HPSv3, PickScore, CLIPScore, ImageReward, Aesthetic, AltCLIP) were also performed on a single H800 GPU.

We include additional qualitative visualizations to complement the quantitative experiments in Section 4. Figure 8 demonstrates SAGE-GRPO's superior ability to align with emotional descriptions in text prompts, capturing subtle facial expressions and mood cues that baselines often miss.

To further validate the effectiveness of different KL strategies discussed in Section 3.2.2, we provide qualitative comparisons across five variants: no KL regularization, Fixed KL (anchored to the initial model π_0), Step-wise KL (velocity control only), Moving KL (position control only), and Dual Moving KL (combining both position and velocity control). Figures 11 and 12 show that Dual Moving KL consistently produces more realistic details, better temporal consistency, and stronger alignment with prompt descriptions compared to other KL strategies, which aligns with the quantitative findings in Section 4.2.

H Ablation on Moving KL Update Frequency N

The hyperparameter N determines the update frequency of the moving KL anchor, meaning the reference model is updated every N step. Therefore N controls the trade-off between the stability



Figure 12: **KL strategy ablation: qualitative comparison (Case 2).** Additional visual comparison demonstrating how different KL strategies affect generation quality. Dual Moving KL consistently achieves better photorealism and alignment with prompt descriptions compared to alternatives, validating the position-velocity control mechanism discussed in Section 3.2.2.

of the trust region constraint and the plasticity needed for continuous reward optimization. We ablate $N \in \{5, 10, 20\}$ under the otherwise-unchanged SAGE-GRPO recipe (HunyuanVideo 1.5, VideoAlign reward, Setting B, 4 nodes, evaluated at step 100) and report the headline metrics in Table 5. N=10 wins on $4/8$ headline metrics, including the most informative VideoAlign axes (Overall, VQ, MQ); N=5 underfits the advantage estimate and over-optimizes simpler texture metrics, while N=20 dilutes the per-step gradient signal and shows the slowest training-reward improvement (ΔR at step 100 is $+0.141$, $+0.442$, $+0.325$ for N=5/10/20 respectively). We therefore use N=10 as the default moving frequency of moving KL.

Table 5: **Ablation on moving frequency N.** All runs use the SAGE-GRPO recipe on HunyuanVideo 1.5 + VideoAlign (Setting B), 4 nodes, evaluated at step 100. **Bold** marks the best entry per metric.

Metric	N=5	N=10	N=20
VideoAlign-Overall ↑	-2.046	$\mathbf{-1.014}$	-1.614
VideoAlign-VQ ↑	-0.832	$\mathbf{-0.292}$	-0.504
VideoAlign-MQ ↑	-0.520	$\mathbf{+0.145}$	-0.313
VideoAlign-TA ↑	$\mathbf{-0.694}$	-0.867	-0.797
Aesthetic ↑	0.074	**0.184**	0.160
CLIPScore ↑	**0.623**	0.579	0.581
PickScore ↑	18.720	18.798	**18.902**
HPSv3 ↑	4.515	5.201	**5.254**
Wins	$2/8$	$\mathbf{4/8}$	$2/8$

I Prompt Setting and Example

Training prompt pool. We maintain a pool of ∼10k long-form, cinematic text prompts that mix realistic, animated, and abstract styles. The pool is generated with Google Gemini 2.5 Pro under a fixed cinematic prompting template that requests concrete physical detail (camera motion, lighting, costume, posture, micro-expression, environmental cues) so that the reward model has enough surface to distinguish small policy changes during GRPO training; the full template, generation seeds, and the released prompt list will be open-sourced together with the code at publication.

Validation prompts (metric and user study). For all metric evaluations and the user study reported in Sections 4.3 and 4.4 we use a single held-out validation pool of 32 prompts, partitioned into three categories: *realistic* (n_p=14 everyday-physics, photo-real prompts; e.g., portraits, urban scenes, nature documentaries), *animated* (n_p=10 stylized prompts; cartoon, anime, claymation, oil-painting), and *abstract* (n_p=8 non-figurative prompts emphasising colour, texture, motion, or ambience without a clear narrative subject). The same 32 prompts are used as the input to all four GRPO variants and to the user study, so that all reported metric scores and pairwise win rates are computed on the same prompt set.

A representative prompt from our public visual gallery is shown below; the full released prompt set covers a similar density of visual and motion cues, and will be released alongside the code at publication.

> "slow, upward, medium shot zooms in on the legendary god of war, Lü Bu, standing in the center of a battlefield strewn with corpses and shrouded in dusk. His ornate beast-headed shoulder armor is shattered, the dull metal riven with deep sword marks and cracks, and several broken arrows remain lodged in the gaps between the armor plates. Beneath his torn, crimson battle robe lies a lining soaked in blood. His signature long hair hangs in tangled strands, matted with sweat and seeping blood, plastered across his handsome face, stained with dust and blood. His eyes, blazing with unyielding rage, stare fixedly ahead. With all his strength, he clings to his equally blood-soaked Fang Tian Hua Ji, digging the heavy weapon deep into the muddy ground beneath his feet to support his shaky body. A tremor, born of extreme exhaustion and blood loss, emanated from deep within his taut muscles, coursing through his entire body, causing his massive frame to tremble uncontrollably. Each heavy breath produced a puff of white mist. The backdrop was an ancient battlefield at dusk. The setting sun, like blood, stained the sky a tragic orange-purple. Thin clouds of dust drifted slowly through the stagnant air. A final ray of light, from behind and side, outlined his stern figure, casting a long, solitary shadow across his tattered armor. The entire scene was filled with epic sadness and unyielding heroism."

J Direct Comparison with TempFlow-GRPO

TempFlow-GRPO (He et al., 2025b) is the closest concurrent alternative to our gradient normalization equalizer (loss-side temporal-flow weighting vs. our marginal-preserving SDE variance $\Sigma_t^{1/2}$). Reproduced under our HunyuanVideo1.5 + VideoAlign framework (Fixed KL), SAGE-GRPO outperforms TempFlow-GRPO on Overall, VQ, TA, and PickScore in both reward settings (Table 6).

Table 6: **SAGE-GRPO vs. TempFlow-GRPO** (HY-Video1.5 + VideoAlign, w/ Fixed KL).

Method	Setting	Overall ↑	VQ ↑	MQ ↑	TA ↑	CLIPScore ↑	PickScore ↑
TempFlow-GRPO	A (averaged)	-1.9799	-0.8242	-0.4700	-0.6857	0.6747	18.6498
SAGE-GRPO	A (averaged)	**0.2244**	**-0.7438**	-0.5320	**1.5001**	0.5446	**19.1932**
TempFlow-GRPO	B (alignment)	-1.9797	-0.7510	-0.5788	-0.6498	0.6708	18.7486
SAGE-GRPO	B (alignment)	**0.2857**	**-0.7062**	**-0.4425**	**1.4344**	0.5414	**19.1802**

K HPSv3 Image Generalization: Per-Prompt Paired t-Tests

We back the SD3.5 / FLUX rows of Table 3 with per-prompt paired t-tests against each baseline (N=768 pairwise comparisons per cell, identical prompts across methods). We focus on four representative metrics—the training reward HPSv3 plus three widely used perceptual scores (PickScore, Aesthetic, AltCLIP)—on which the relative ordering of SAGE-GRPO and the baselines is most diagnostic. Δ = SAGE-GRPO − baseline, and win-% is the fraction of prompts on which SAGE-GRPO scores higher.

Across the six HPSv3 head-to-heads (3 backbones × 2 strongest baselines, dropping FlowGRPO on SD3.5 where it collapses and on FLUX where it is the weakest), SAGE-GRPO records 3 significant wins (both head-to-heads on HunyuanVideo 1.5 with $\Delta > 0.7$, and SD3.5 vs. DanceGRPO at $p = 0.001$), 3 statistical ties (SD3.5 vs. CPS at p=0.13; FLUX vs. DanceGRPO at p=0.30; FLUX vs. CPS at p=0.060), and *zero significant losses* on the training reward—a property no other method achieves.

Table 7: **HPSv3 paired t-tests on SD3.5-medium** (training reward = HPSv3, N=768 per cell, evaluated at training steps $\{1440, 1500, 1560\}$). Δ = SAGE-GRPO − baseline; positive entries favor SAGE-GRPO, win-% is the fraction of prompts on which SAGE-GRPO scores higher.

Metric	**vs. DanceGRPO** Δ (win-%)	**vs. FlowGRPO** Δ (win-%)	**vs. CPS** Δ (win-%)
HPSv3 ↑	+0.139 (58.9%)	+0.960 (78.1%)	−0.062 (47.0%)
PickScore ↑	+0.078 (58.3%)	+0.260 (66.3%)	−0.000 (51.2%)
Aesthetic ↑	+0.000 (52.1%)	+0.056 (58.5%)	−0.009 (46.9%)
AltCLIP ↑	+0.142 (45.2%)	+0.292 (52.3%)	−0.017 (44.5%)

Table 8: **HPSv3 paired t-tests on FLUX.1-dev** (training reward = HPSv3, N=768 per cell, evaluated at training steps $\{1440, 1500, 1560\}$). Same conventions as Table 7. The four methods cluster tightly on this backbone; SAGE-GRPO ties or beats every baseline on HPSv3 with no significant loss.

Metric	**vs. DanceGRPO** Δ (win-%)	**vs. FlowGRPO** Δ (win-%)	**vs. CPS** Δ (win-%)
HPSv3 ↑	−0.023 (47.0%)	+0.105 (62.1%)	+0.039 (51.0%)
PickScore ↑	−0.026 (39.1%)	+0.052 (57.9%)	+0.000 (50.1%)
Aesthetic ↑	−0.032 (38.9%)	−0.003 (51.6%)	−0.021 (41.0%)
AltCLIP ↑	−0.105 (32.6%)	−0.027 (34.2%)	+0.153 (53.5%)

L LIMITATIONS: THE REWARD-MODEL BOTTLENECK

The most important caveat of all our results, and of video GRPO as a whole, is that the alignment quality is upper-bounded by the reward model used to drive the optimization. We use VideoAlign (Liu et al., 2025c) as the video reward model, which is biased toward photorealistic perceptual quality and is prone to reward hacking on its own VQ/MQ axes (Li et al., 2025b); and HPSv3 (Ma et al., 2025b), a strong frame-level human preference scorer that is designed for images rather than video and therefore cannot reliably score motion quality, temporal consistency, or long-range causal coherence. Neither model captures cinematographic constructs such as camera kinematics, scene composition, narrative pacing, or multi-shot coherence, all of which our human evaluators clearly weigh in the user study. As a consequence, even a perfectly converged GRPO policy on these rewards inherits the reward model's blind spots: stylized prompts collapse toward photo-realism, motion smoothness can be over-optimized at the expense of action expressiveness, and any improvement we observe on HPSv3 transfers to human preference only to the extent that HPSv3 itself does. We therefore view the development of stronger video reward models, jointly trained on temporal coherence, motion quality, semantic alignment, and stylistic faithfulness, as the single largest open problem for video GRPO going forward. SAGE-GRPO is intentionally reward-model agnostic and can be re-applied to such future rewards without modification of the optimization framework.