ONE LENGTH DOES NOT FIT ALL: UNIFYING VIDEO RL ACROSS LENGTH SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Following its success in aligning large language models with human preferences, reinforcement learning (RL) post-training has become a dominant paradigm for aligning video generation models with user preferences. However, existing approaches typically perform RL on a single, fixed video length: restricted to a single length scale, the learned improvements amount to local adaptations that fail to generalize across durations. This raises a fundamental question unexplored: how do the optimization dynamics of RL inherently differ across video length scales? In this work, we systematically reveal a severe **Length-Scale Inconsistency** in video RL. Since short and long videos place different emphasis on information dimensions during optimization, training under the same reward function over-exploits whichever dimensions dominate the reward at that length: static color and texture cues for short videos, whereas temporal shortcuts cause incoherent motion and abrupt transitions for long videos. To address this, we propose the Length-Aware Training Framework (LA-Framework), which unifies RL for videos of varying lengths through two core designs that directly resolve the underlying causes: (1) a length-conditioned timestep credit assignment mechanism that efficiently estimates reward sensitivity via *sparse branching* and *shared prefixes*, constructing a weight tree to dynamically allocate gradient budgets to the timesteps that actually determine the reward at each length; and (2) a multi-length joint training paradigm that normalizes advantages within identical length groups, allowing length-specific hacking directions to counteract each other while jointly reinforcing genuinely beneficial improvements. Extensive experiments across nine evaluation dimensions demonstrate that our framework effectively mitigates length-specific reward hacking and achieves high generation quality consistently across various video lengths.

1 INTRODUCTION

Video generation models (Brooks et al., 2024; Yang et al., 2025; Wan et al., 2025) synthesize dynamic visual content by progressively denoising random noise through diffusion or flow matching (Ho et al., 2020; Lipman et al., 2022; Savani et al., 2026) processes. As a highly promising post-training paradigm, reinforcement learning (RL) was initially proposed as RLHF to align large language models with human preferences (Ouyang et al., 2022; Shao et al., 2024), and has been widely adopted to align video generation models with human preferences, providing distinct advantages in three main aspects: 1) By directly optimizing a reward model that captures human aesthetic and semantic preferences, RL significantly enhances the visual quality and text-video alignment of the generated content (Liu et al., 2026b). 2) Unlike supervised fine-tuning (Wei et al., 2021; Ouyang et al., 2022) that requires massive paired datasets, RL can effectively improve model performance using reward signals alone, making it a scalable optimization method (Kaufmann et al., 2023). 3) The integration of RL with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has shown great potential in stabilizing training and reducing the variance of policy gradients in continuous generation tasks. Extensive efforts have been made to improve video generation models via reinforcement learning (Zhang et al., 2026; Liu et al., 2026a; Xue et al., 2025). However, these methods typically conduct training on a single, fixed video length (He et al., 2026; Liu et al., 2026b). Under such fixed-length optimization, the learned improvements are tailored to the statistics of a specific duration, and may introduce length-specific biases that degrade generation quality elsewhere. This raises a fundamental yet underexplored question: *how do the optimization dynamics and effects of reinforcement learning inherently differ when applied to short clips versus long videos?*

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.

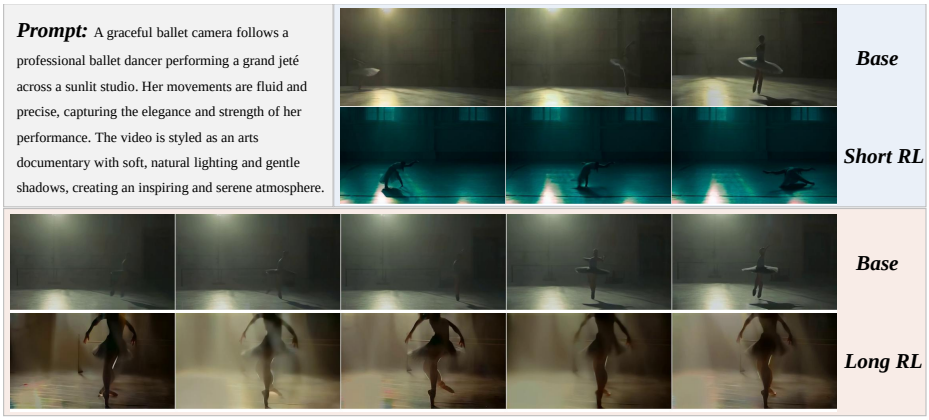


Figure 1: Fixed-length RL training exhibits distinct behaviors across video lengths under the same reward: static attribute exploitation for short videos versus temporal inconsistency for long videos.

To answer this question, we conduct systematic experiments and reveal a severe **Length-Scale Inconsistency** in video reinforcement learning. As shown in Fig. 1, trained with the same reward function, fixed-length RL exhibits distinct reward hacking (Skalse et al., 2022; Bondarenko et al., 2025) behaviors across video lengths: *short-video training tends to exploit static visual attributes such as color and texture to deceive the reward model, while long-video training drives the policy to exploit temporal shortcuts, resulting in inconsistent motion and discontinuous transitions.* Consequently, improvements achieved on a single length cannot be trivially extended to others, highlighting the need for a unified training paradigm across length scales.

In this work, we propose the Length-Aware Training Framework (LA-Framework), which unifies the reinforcement learning training paradigm for both short and long videos. Our approach is built on two core pillars. **1) Length-conditioned timestep credit assignment.** Existing methods for video alignment uniformly distribute the terminal reward across all timesteps, ignoring their distinct contributions to the final reward, which misallocates the optimization budget and exacerbates length-specific hacking. Although non-uniform credit assignment has been actively explored by GRPO variants in LLM post-training, it remains absent from length-heterogeneous video generation. Instead, LA-Framework adaptively allocates the optimization budget according to the reward sensitivity of each timestep under a specific video length, so that the long-range motion and cross-frame consistency required by long videos and the fine-grained visual quality pursued by short videos are optimized in a coordinated manner. **2) Multi-length joint training.** Training on a single fixed length encourages the model to overfit length-specific generation strategies. However, naively pooling different lengths into one training process is non-trivial: reward statistics differ drastically across lengths, and directly pooling advantages would bias the policy toward the length with the largest gradient magnitude. We therefore jointly train on multiple video lengths of the same prompt and normalize advantages within identical length groups, which prevents gradient domination by any single length and, more importantly, forces length-specific hacking directions to counteract each other. We further balance the sampling ratio across lengths to keep the optimization pressure even, enabling the model to acquire universal generation strategies across length scales and offering a practical path toward unified RL optimization for videos of diverse durations.

Concretely, LA-Framework instantiates the credit assignment in two cascaded steps, *prior-guided allocation* followed by *sensitivity-driven refinement*: ***Step 1: prior-guided allocation.*** Since the long-range motion and cross-frame consistency critical for long videos are largely determined at the early high-noise stages, we introduce a length-conditioned prior $w_{\text{prior}}(t, L)$ that allocates larger optimization budgets to these stages for long videos. As a static, hand-crafted design, however, this prior captures only length-level regularities and remains blind to trajectory-specific variations, motivating a data-driven refinement. ***Step 2: sensitivity-driven refinement.*** We then refine the prior allocation with trajectory-level evidence. In video RL, the reward is available only after the entire video is generated, offering no step-level supervision. We thus probe each timestep by branching B candidate continuations from its shared prefix and define the *reward sensitivity* $S(t, L)$ as the standard deviation of their terminal rewards (Eq. 3), a step-level quantity derivable directly from the terminal reward. Since each branch entails a costly denoising rollout, exhaustively branching at all T timesteps is prohibitive. We instead sparsely branch at M representative timesteps with shared

prefixes. Fusing $S(t, L)$ with $w_{\text{prior}}(t, L)$ and normalizing yields the final timestep weight $W(t, L)$, completing a coarse-to-fine hierarchical allocation and producing the length-conditioned weight tree that dynamically redistributes the gradient budget. The model is then updated with a multi-length joint training paradigm, where prompts are sampled across a length pool, and advantages are normalized within specific lengths to avoid optimization conflicts, forcing the model to learn consistent generation strategies that resist length-specific hacking.

Extensive experiments validate the existence of Length-Scale Inconsistency in video reinforcement learning, revealing that fixed-length training induces distinct length-specific reward hacking behaviors. Moreover, we show that LA-Framework, through multi-length joint training, effectively mitigates these hacking patterns and enables unified optimization across length scales. Finally, comprehensive evaluations on standard video generation benchmarks demonstrate that, compared with existing RL methods, our approach achieves highly competitive generation quality across various video lengths. Our contributions can be summarized as follows:

- We systematically identify and address the Length-Scale Inconsistency in video reinforcement learning, providing a new perspective for optimizing video generation models across different lengths;
- We propose Length-Aware Training Framework (LA-Framework), which efficiently estimates timestep reward sensitivity via sparse branching and shared prefixes, and constructs a length-conditioned weight tree to dynamically allocate gradient budgets;
- Extensive experiments across nine evaluation dimensions demonstrate the effectiveness of our method. Specifically, the framework promotes consistent reinforcement learning optimization across different video length scales and effectively mitigates length-specific reward hacking.

2 Related Work

Reinforcement Learning for Diffusion Generation. Reinforcement learning has emerged as a dominant paradigm for aligning diffusion generation models with human preferences (Li et al., 2026; Black et al., 2024; Xu et al., 2023; Wallace et al., 2024; Wu et al., 2026; He et al., 2026). DDPO (Black et al., 2024) pioneered this by formulating the denoising process as a Markov decision process. ReFL (Xu et al., 2023) enables direct gradient backpropagation through differentiable reward functions. DPOK (Fan et al., 2023) introduces KL regularization to stabilize training dynamics. These techniques have been extended to video generation (Prabhudesai et al., 2024; Liu et al., 2026b; Abbasi et al., 2025; Liu et al., 2026c; Wang et al., 2026b). VADER (Prabhudesai et al., 2024) and VideoReward (Liu et al., 2026b) apply reward-guided optimization to text-to-video diffusion models. DanceGRPO (Xue et al., 2025) and Flow-GRPO (Liu et al., 2026a) adapt generalized policy optimization to flow matching architectures. Despite these advances, existing methods typically train on a single fixed video length. They distribute the final video-level reward uniformly across all generation timesteps, ignoring the underlying credit assignment. Standard policy gradient methods often suffer from this misalignment when dealing with sparse rewards (Sutton et al., 1998; Shao et al., 2024; Schulman et al., 2017; Lightman et al., 2024). In contrast to existing methods that optimize a single length scale, our work focuses on unified reinforcement learning across varying video lengths to address this misalignment.

Reward Hacking and Over-optimization. The phenomenon of reward over-optimization, where proxy reward models are exploited during fine-tuning, has been systematically studied (Wang et al., 2026a; Gao et al., 2023; Yu et al., 2025; Skalse et al., 2022; Casper et al., 2023). (Gao et al., 2023) established scaling laws showing that true quality decays as the policy diverges from the reference policy. (Black et al., 2024) observed this in image generation, where VLM based proxy rewards were exploited through stylistic shortcuts, leading to a collapse in generation diversity. Subsequent efforts proposed mitigation strategies like reward model ensembling (Ramé et al., 2024) and mutual information regularization (Miao et al., 2024; Eisenstein et al., 2023). However, these studies focus exclusively on single-frame. In video generation, reward hacking exhibits distinct length-specific patterns. Short videos tend to exploit static visual attributes like color and texture to deceive the reward model, while long videos are more susceptible to temporal artifacts such as inconsistent motion and discontinuous transitions. Unlike prior work focusing on single-frame, our work investigates the unique hacking behaviors that emerge across different video length scales.

3 METHOD

In this section, we present **Length-Aware Training Framework (LA-Framework)**, a length-conditioned tree-structured optimization framework that unifies the reinforcement learning training paradigm for both short and long videos. We first formulate the problem of length-specific overfitting in fixed-length RL (Sec. 3.1). We then introduce our core length-conditioned timestep credit assignment mechanism, which pre-allocates optimization budgets via a length-conditioned prior and refines them with reward sensitivity estimated via *sparse branching* and *shared prefixes* (Sec. 3.2). Finally, we present the multi-length joint training paradigm and the overall optimization objective (Sec. 3.3). An overview of the proposed framework is illustrated in Fig. 2.

3.1 PRELIMINARIES AND PROBLEM FORMULATION

Given an initial noise $x_1 \sim \mathcal{N}(0, I)$, a video generation model produces the final video x_0 through step-by-step denoising, and we denote the generation trajectory as $\tau = \{x_1, x_{1-\Delta t}, \ldots, x_0\}$. Under the GRPO framework, for a prompt c, the model samples a group of G candidate videos $\{x_0^{(i)}\}_{i=1}^{G}$, and a reward model $R(\cdot)$ computes $r_i = R(x_0^{(i)}, c)$ for each candidate. The group-relative advantage is:

$$A_i = \frac{r_i - \mu(r)}{\sigma(r) + \epsilon}. \tag{1}$$

where $\mu(r)$ and $\sigma(r)$ are the mean and standard deviation of the reward within group G, respectively. Notably, videos of different target lengths share the same number of generation timesteps. Our focus is not the number of denoising steps, but how the reward sensitivity of each timestep varies with the target length within this shared timestep space.

Existing video RL methods typically train on a fixed length L_0. Since the reward model provides only sparse video-level feedback, the model easily overfits the reward distribution and its biases at this specific length, and evolves into reward hacking when exploitable biases exist, exploiting local visual patterns or temporal shortcuts to obtain high rewards. As these strategies are tied to a specific length scale, performance on other video lengths degrades significantly, creating a generalization gap between training and testing lengths. We term this problem **Length-Scale Inconsistency**.

A key factor that amplifies this inconsistency is the **coarse credit assignment** of standard video RL (Liu et al., 2026a). The advantage A_i in Eq. equation 1 is a terminal reward obtained only after the entire video is generated, yet it is uniformly broadcast to all timesteps, i.e., $\tilde{A}_{i,t} = A_i$ for $t \in \mathcal{T}$, implicitly assuming equal reward sensitivity across timesteps. However, different generation timesteps do not possess equal reward sensitivity:

- **High-noise stages** ($t \to 1$): determine global structure and coarse motion trajectories.
- **Low-noise stages** ($t \to 0$): handle local details and texture refinement.

More importantly, this importance *shifts* with video length: long videos demand complex long-range motion and cross-frame semantic consistency, while short videos have simpler structures where local details may dominate the final reward. The contribution of a timestep t is thus a length-dependent function: $C(t, L) = \mathrm{Credit}(t \mid L)$. Motivated by this, LA-Framework explicitly estimates $C(t, L)$ via a two-stage, tree-structured credit assignment, as illustrated in Fig. 2.

3.2 CREDIT ASSIGNMENT TREE VIA SPARSE BRANCHING AND SHARED PREFIXES

To address the coarse credit assignment analyzed in Sec. 3.1, we instantiate the length-conditioned timestep credit assignment in two cascaded steps: a *prior-guided allocation* that pre-assigns budgets from length-level regularities, followed by a *sensitivity-driven refinement* that calibrates them with trajectory-level reward feedback (middle part of Fig. 2).

Step 1: prior-guided allocation. We introduce a length-conditioned prior that pre-allocates the optimization budget across timesteps:

$$w_{\mathrm{prior}}(t, L) = |T| \cdot \frac{\exp\left(\beta \cdot \frac{L}{L_{\max}} \cdot t\right)}{\sum_{\tau \in \mathcal{T}} \exp\left(\beta \cdot \frac{L}{L_{\max}} \cdot \tau\right)}, \tag{2}$$

where $|T|$ is the number of timesteps, $L_{\max}$ is the maximum length in the training pool, and $\beta > 0$ is a temperature controlling the prior. The prior satisfies $\sum_{t \in \mathcal{T}} w_{\mathrm{prior}}(t, L) = |T|$, keeping the total update budget consistent across lengths: it approaches a uniform distribution for short videos,

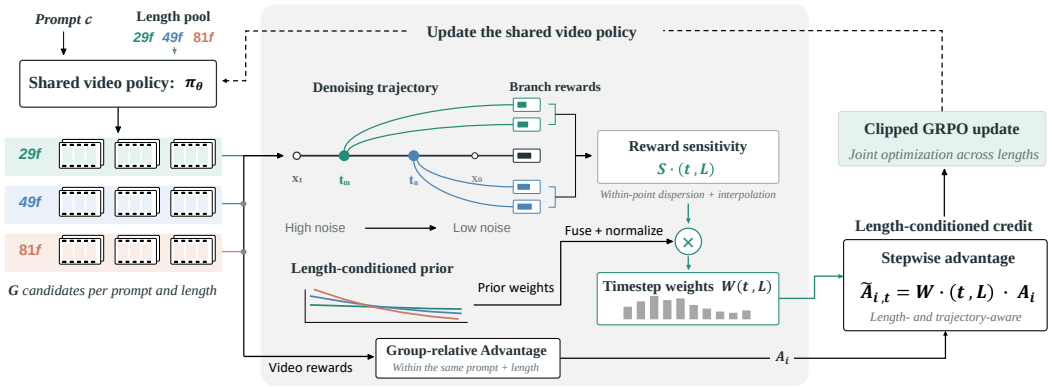


Figure 2: **Overview of Length-Aware Training Framework (LA-Framework).** Given a prompt and a length pool, the shared video policy π_θ samples G candidates per prompt–length pair, denoised over the same T timesteps. **Middle:** Sparse branching rollouts at selected timesteps t_m, whose branch rewards (sharing the full prefix) measure the reward sensitivity $S(t, L)$. The sensitivity is fused with a length-conditioned prior to build the timestep weights $W(t, L)$, forming a length aware credit assignment tree. **Right:** the group-relative advantage is modulated into the stepwise advantage for the clipped GRPO update, jointly optimized across all lengths.

whereas for larger L it gradually shifts toward early high-noise stages, where long-range motion and cross-frame consistency are largely determined, reserving budget for global structure and long-range motion planning. As a static, hand-crafted design, however, this prior captures only length-level regularities and remains blind to trajectory-specific variations, motivating a data-driven refinement.

Step 2: sensitivity-driven refinement. We then refine the prior with **sparse branching** over **shared prefixes**. Since full branching at all T timesteps is prohibitive for video generation models, we uniformly select M representative timesteps as branch points and generate B candidate continuations from the state x_{t_m} of each calibration trajectory. Crucially, these B branches share the complete prefix from the initial noise x_1 to x_{t_m} and require no regeneration, reducing the rollout overhead from $T \times B$ to $M \times B$ (Li et al., 2026). Each branch continues the diffusion process from x_{t_m} to completion, yielding branch rewards $\{R^{(b)}_{t_m,L}\}_{b=1}^{B}$. If branches sharing the same state lead to substantially different final rewards, the decisions at this timestep strongly affect the final quality. We therefore measure the reward sensitivity as the standard deviation of branch rewards:

$$S(t_m, L) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(R^{(b)}_{t_m,L} - \bar{R}_{t_m,L}\right)^2}, \tag{3}$$

where $\bar{R}_{t_m,L}$ is the mean of the branch rewards. A larger $S(t_m, L)$ marks a more decisive timestep that deserves a larger update budget. For non-branch timesteps, $S(t, L)$ is obtained via linear interpolation, yielding a complete sensitivity profile. We normalize the sensitivity as $\hat{S}(t, L) = \frac{S(t,L)}{\frac{1}{|T|}\sum_{\tau\in\mathcal{T}} S(\tau,L)+\epsilon}$ and fuse it with the prior (Eq. equation 2):

$$W(t, L) = |T| \cdot \frac{w_{\text{prior}}(t, L)\hat{S}(t, L)}{\sum_{\tau\in\mathcal{T}} w_{\text{prior}}(\tau, L)\hat{S}(\tau, L)}. \tag{4}$$

This guarantees $\sum_{t\in\mathcal{T}} W(t, L) = |T|$ with two properties: (1) different video lengths automatically induce different timestep update distributions; (2) the weights are calibrated once on a held-out set of trajectories before training and frozen during training. Since calibration is a standalone inference-time procedure, this freezing is readily relaxed: the weights can be periodically re-estimated every R training steps ($R{=}\infty$ in our default configuration) to track the evolving policy. As shown in Fig. 2, this yields a credit assignment tree jointly determined by the video length and the branch reward feedback. We finally replace the uniform timestep advantage with the length-conditioned stepwise advantage:

$$\tilde{A}_{i,t} = W(t, L) \cdot A_i, \tag{5}$$

where A_i is the group-relative advantage in Eq. equation 1. In this way, LA-Framework dynamically adjusts the gradient budget of each timestep according to the video length, concentrating the update on the generation stages that actually contribute to the current length scale.

3.3 Multi-Length Joint Training Paradigm and Optimization Objective

A single length-conditioned credit assignment still trains one length at a time and cannot fully resolve the cross-length generalization gap. To unify optimization across length scales, we propose a **multi-length joint training paradigm**. As depicted in Fig. 2, we maintain a length pool $\mathcal{L}_{pool} = \{L_1, L_2, \ldots, L_K\}$, and for each prompt c_j in a batch $\mathcal{B}$, the model samples G candidate videos for every length $L_k \in \mathcal{L}_{pool}$, all sharing the same $T = 20$ timesteps while using their own length-conditioned credit weights via Sec. 3.2.

Different video lengths exhibit different visual complexities and reward distributions, normalizing advantages directly over a joint batch would misinterpret inherent length-wise reward differences as generation quality differences. We therefore normalize advantages only within same-prompt, same-length candidate groups:

$$A_{j,k,i} = \frac{R(x_0^{(j,k,i)}, c_j) - \mu_{j,k}}{\sigma_{j,k} + \epsilon}. \tag{6}$$

With this length-conditioned intra-group normalization, lengths compare relative generation quality within their own candidate spaces instead of competing for absolute reward magnitudes, avoiding optimization conflicts caused by length distribution differences.

Combining the credit weights $W(t, L_k)$ in Eq. equation 4 with the group advantage $A_{j,k,i}$ in Eq. equation 6 yields the final stepwise advantage $\tilde{A}_{j,k,i,t} = W(t, L_k) \cdot A_{j,k,i}$, which generalizes Eq. equation 5 to the multi-length setting. The model is then updated with a clipped policy objective based on importance sampling:

$$\mathcal{L}_{\text{policy}}(\theta) = -\mathbb{E}_{j\sim\mathcal{B},\, k\sim\mathcal{L}_{pool},\, i\sim\mathcal{G},\, t\sim\mathcal{T}} \left[\min \left(\rho_{j,k,i,t} \tilde{A}_{j,k,i,t},\ \text{clip}\left(\rho_{j,k,i,t}, 1-\epsilon, 1+\epsilon\right) \tilde{A}_{j,k,i,t} \right) \right], \tag{7}$$

where $\rho_{j,k,i,t} = \frac{\pi_\theta\left(x_{t-\Delta t}^{(j,k,i)} | x_t^{(j,k,i)}, c_j\right)}{\pi_{\theta_{\text{old}}}\left(x_{t-\Delta t}^{(j,k,i)} | x_t^{(j,k,i)}, c_j\right)}$ is the importance sampling ratio.

Through the design of Eq. equation 7, LA-Framework introduces multiple video lengths simultaneously in a single training process and establishes an independent timestep credit assignment structure for each length. Videos of different lengths receive different timestep update weights, forming the length-conditioned update weight tree illustrated in Fig. 2. The model no longer optimizes rewards for a single fixed video length, but learns simultaneously across different length scales: for long videos, it allocates more optimization budget to the generation stages sensitive to long-range structure and temporal consistency; for short videos, it maintains a more balanced timestep update based on actual reward feedback. Through the synergy of multi-length sampling (Sec. 3.3), length-conditioned credit assignment (Eq. equation 4), and intra-group advantage normalization (Eq. equation 6), LA-Framework effectively mitigates length-specific reward hacking and promotes consistent reinforcement learning optimization across different video length scales. The details of this algorithm process are in Appendix A.

4 Experiments

4.1 Experimental Setup.

Models: We conduct experiments on two representative video generation backbones with different architectures and scales: Wan2.1-1.3B (Wan et al., 2025) and CogVideoX-2B (Yang et al., 2025). To investigate the optimization dynamics across length scales, our training uses a length pool of {29, 49, 81} frames; the evaluation lengths 33 and 57 frames are deliberately excluded from the pool and held out to assess cross-length generalization.

To comprehensively measure different quality dimensions, we employ five widely-adopted proxy rewards. For static visual quality, we adopt CLIPScore (Hessel et al., 2021) for text-image alignment, PickScore (Kirstain et al., 2023) and HPSv2 (Wu et al., 2023) for human preference alignment, and the LAION Aesthetic predictor for visual aesthetics. These image-based rewards are computed frame-wise and averaged over the whole video. For video-native rewards, we adopt VideoAlign (Liu et al., 2026b), a VLM-based multi-dimensional reward model that jointly evaluates visual quality, motion quality, and text alignment.

Evaluation: We evaluate generation quality on two widely-used video generation benchmarks: VBench (Huang et al., 2024) and VideoScore (He et al., 2024). VBench decomposes video generation quality into multiple hierarchical dimensions, which allows fine-grained analysis of both

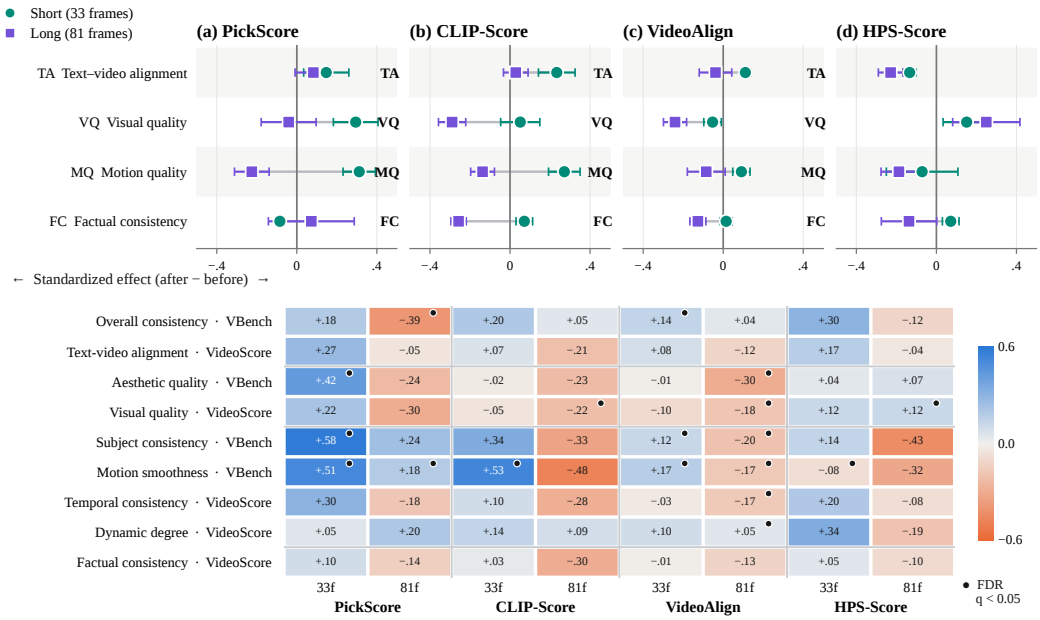


Figure 3: **Standardized effects of fixed-length RL.** For each proxy reward, we report the quality change (after − before) of short and long training over four dimensions: text–video alignment (TA), visual quality (VQ), motion quality (MQ), and factual consistency (FC). Fixed-length RL exhibits clearly divergent and inconsistent optimization effects across video lengths.

static and temporal quality. VideoScore is an automatic evaluator that simulates fine-grained human feedback, providing scores across dimensions such as text–video alignment, visual quality, motion smoothness, temporal consistency, and dynamic degree. We compare our LA-Framework against several representative reinforcement learning approaches for visual generation, including Flow-GRPO (Liu et al., 2026a), DanceGRPO (Xue et al., 2025), and DDRL (Ye et al., 2025).

4.2 Empirical Observations: Length-Scale Inconsistency and Length-Specific Reward Hacking.

In this section, we examine how fixed-length RL behaves across video length scales. Figure 3 reports the standardized change in true generation quality (after − before) over four dimensions: **text-video alignment (TA), visual quality (VQ), motion quality (MQ), and factual consistency (FC)**. A clear divergence emerges across scales: trained with identical rewards, short-video RL and long-video RL drive quality changes along fundamentally different directions.

Training on short videos mainly improves static visual attributes. Short-video RL yields large VQ gains across proxy rewards, accompanied by moderate TA improvements, while motion-related metrics remain flat or even negative, indicating a mild reward hacking at the static level: the policy tends to enrich colors, contrast, and textures favored by proxy rewards. As visualized in Figure 1 (top row), the short-video-RL model produces visibly richer colors and finer textures than the base model, yet its motion becomes stiffer and more repetitive. As the training length increases, this pattern reverses with video length: for long videos, the VQ gains shrink or even reverse, while MQ-related metrics degrade severely, motion smoothness drops by up to 0.48 under CLIP-Score, and overall consistency drops by up to 0.39 under PickScore. Correspondingly, in Figure 1 (bottom row), the long-video-RL model preserves a high-quality prefix but exhibits inconsistent motion and abrupt transitions in later frames (extended visualizations and per-metric analyses in Appendix B). Since neither pattern appears in the base model, these artifacts are induced by length-specific over-optimization rather than inherited from it. Overall, such length-divergent hacking is systematic across all five proxy rewards: the hacking direction is itself a function of the training length. We term this phenomenon the **Length-Scale Inconsistency**, which motivates our length-aware training framework.

4.3 Effectiveness in Unified Training of Length-Aware Training Framework

In this section, we show that the LA-Framework unifies reinforcement learning optimization across video length scales. Figure 4 visualizes the cross length consistency of the optimization effects

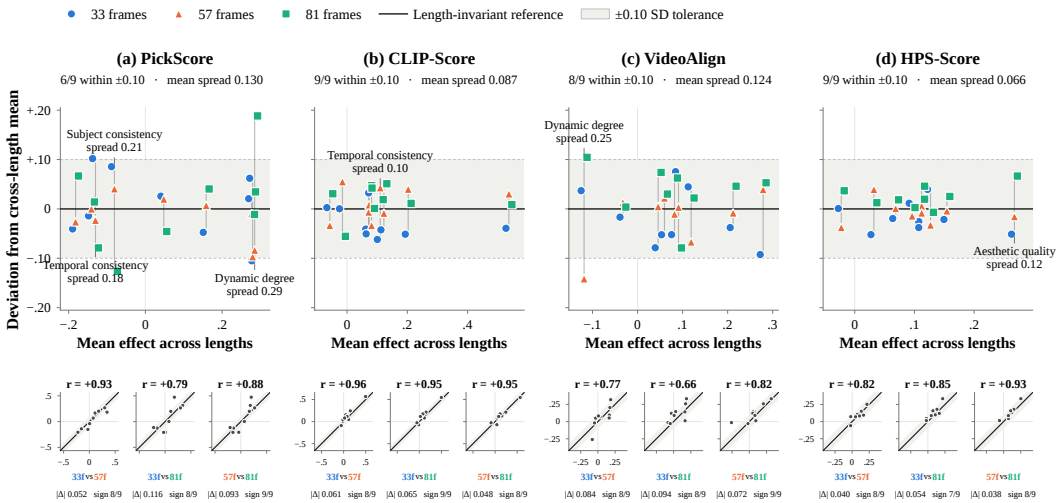


Figure 4: **Cross-length consistency of optimization effects under unified training.** For each proxy reward, the top panel plots the mean standardized effect of each evaluation dimension across lengths (x-axis) against its deviation from the cross-length mean (y-axis), for 33-, 57-, and 81-frame videos. The black line denotes the length-invariant reference, and the shaded band indicates the ± 0.10 SD tolerance. The bottom panel shows the pairwise correlations of per-dimension effects between length pairs, reported with Pearson correlation r, mean absolute deviation $|\Delta|$, and sign agreement. Optimization effects are consistent across video lengths for all proxy rewards.

under our joint-length training: for each proxy reward, the top panel plots the mean standardized effect of each evaluation dimension across lengths against its deviation from the cross length mean; the bottom panel reports the pairwise correlations of per-dimension effects between length pairs.

Optimization effects are consistent across lengths. Under unified training, the effects of each evaluation dimension remain tightly clustered around the length-invariant reference: the mean cross-length spread ranges from only 0.066 (HPS-Score) to 0.130 (PickScore), and 32 out of 36 dimension-reward pairs fall within the ± 0.10 SD tolerance. The few exceptions (e.g., dynamic degree and subject consistency) remain bounded. Notably, the mean effects are also predominantly positive. Since length-specific hacking would instead manifest as inflated yet divergent effects on specific dimensions (Sec. 4.2), this pattern suggests that the LA-Framework effectively suppresses reward hacking: when videos of different lengths are jointly trained, their length-specific hacking directions (e.g., static exploitation for short videos versus temporal shortcuts for long videos) mutually counteract, whereas genuine improvements, which are shared across lengths, are jointly reinforced. Consequently, only length-invariant optimization directions survive, yielding consistently positive effects across all lengths. The bottom panel of Figure 4 further shows strong pairwise agreement between lengths, with Pearson correlations ranging from $+0.66$ to $+0.96$ across all reward and pair combinations. This indicates that the model acquires a shared set of optimization directions that generalize across length scales, rather than length-specific behaviors. In contrast to fixed-length RL, unified training with LA-Framework yields highly consistent effects across all lengths, including the two evaluation lengths unseen during training and proxy rewards (complete underlying results in Appendix C). These results verify that LA-Framework unifies optimization across length scales and suppresses length-specific reward hacking.

4.4 From Cross-Length Consistency to Genuine Quality.

To compare fairly with existing RL methods, we instantiate $R(\cdot)$ with VideoAlign and apply LA-Framework to two models. Since fixed-length training exhibits length-specific effects (Sec. 4.2), evaluating at a single length would favor each method at its own training length. We therefore evaluate all methods at both a short (29-frame) and a long (49-frame) length and report the average, with baselines trained at their typical fixed lengths and ours retaining multi-length training (Appendix A).

RL improves most metrics over the base models for every method in Table 1, confirming the general effectiveness of RL for video generation. The comparison is therefore not about whether RL helps, but about which method converts the reward signal into genuine quality most faithfully. In

Table 1: Comparison with representative video RL methods. All methods are trained with the VideoAlign reward, and scores are averaged over the short (29-frame) and long (49-frame) evaluation lengths. Best in **bold**, second best underlined. LA-Framework ranks first on 17 of 18 metrics across both base models without sacrificing any dimension, showing that the benefits of unified optimization and hacking mitigation translate into genuinely higher generation quality.

Model	Method	VBench ↑				VideoScore ↑				
		Subject consistency	Aesthetic quality	Motion smoothness	Imaging Quality	VQ	TC	DD	TVA	FC
Wan2.1	Baseline	86.88	45.90	91.01	45.93	71.44	77.10	65.89	68.07	57.67
	DanceGRPO	90.17	45.30	92.42	45.11	73.33	77.54	68.74	68.39	62.60
	FlowGRPO	90.20	44.46	92.17	44.67	74.00	78.63	70.44	70.81	64.88
	DDRL	88.14	46.88	91.83	46.13	73.93	79.98	69.51	69.28	66.17
	Ours	**93.58**	**50.50**	**94.33**	**46.89**	**75.83**	**80.81**	**71.30**	**71.47**	**67.10**
CogVideoX	Baseline	85.14	42.08	88.73	37.86	63.71	74.23	64.28	63.55	53.33
	DanceGRPO	87.19	41.09	88.80	40.20	65.58	72.87	64.94	66.83	55.51
	FlowGRPO	85.86	43.26	90.75	39.82	66.63	74.15	66.35	65.05	58.52
	DDRL	88.75	44.83	89.64	40.01	66.07	**77.60**	65.77	67.59	56.44
	Ours	**90.46**	**45.61**	**91.22**	**41.99**	**67.52**	77.24	**68.23**	**68.51**	**59.87**

this regard, averaging over the two lengths is revealing: the baselines' length-specific gains do not transfer, as an improvement on one length is offset on the other, and their gains concentrate on the VideoScore dimensions most related to the VideoAlign reward itself, while the independent VBench evaluation exposes collateral degradations. This reproduces the reward hacking signature identified in Sec. 4.2 under the fixed-length setting.

As shown in Table 1, our method ranks first on all nine metrics, with no dimension sacrificed relative to the base models. Its gains are uniform across the two evaluation lengths rather than traded off between them, and the largest margins appear exactly on the dimensions most vulnerable to fixed-length RL(dynamic degree, factual consistency), where gains on one length typically come at the cost of the other. These results indicate that the benefits of LA-Framework are not confined to ***cross-length consistency and hacking mitigation***: by assigning the optimization budget according to the actual reward sensitivity of each length, the framework converts the conflicting, length-specific gains of fixed-length RL into coherent and length-stable quality improvements. That our scores remain highest on nearly every metric after averaging directly corroborates the unified consistency of Sec. 4.3: gains that hold on both lengths, rather than being traded off between them, accumulate into genuinely higher generation quality (visualization in Appendix E).

We further conduct ablation studies and efficiency analysis in Appendix D. The ablations confirm that sparse branching, the length-conditioned prior, and intra-group normalization each play a distinct and non-redundant role, while the efficiency analysis shows that the additional training cost is modest and inference remains overhead-free.

5 Conclusion

In this paper, we investigate the optimization dynamics of reinforcement learning across video lengths and reveal a severe **Length-Scale Inconsistency**: trained with the same reward, fixed-length RL induces length-specific reward hacking. We propose the **Length-Aware Training Framework**, which unifies RL across lengths through two designs: (1) a length-conditioned timestep credit assignment mechanism, which estimates the reward sensitivity of each timestep via sparse branching with shared prefixes and dynamically reallocates the gradient budget through a weight tree; and (2) a multi-length joint training paradigm, which normalizes advantages within identical length groups so that length-specific hacking directions counteract each other while genuine improvements are jointly reinforced. Extensive experiments show that LA-Framework effectively mitigates length-specific reward hacking and achieves consistent, high-quality generation across video lengths, pointing toward a unified RL post-training paradigm for video generation at diverse durations.

AI USE STATEMENT

We use large language models (LLMs) in two limited aspects: polishing the writing and checking grammatical accuracy of this manuscript, and assisting in verifying the experiment code and cross-checking the reported results. All research ideas, method designs, and experiments are conducted solely by the authors, without any further use of LLMs beyond the above scope.

ETHICS & REPRODUCIBILITY STATEMENT

In this work, we study reinforcement learning post-training for video generation across different video lengths, which imposes no negative social impacts: all base models, reward models, and evaluation benchmarks are publicly available, and no human subjects or private data are involved. Moreover, our method only realigns generation models with human preferences and introduces no new risks beyond those of the underlying models, whose licenses we strictly follow. Detailed experimental configurations are provided in the "Experimental Settings" section to ensure reproducibility. Experiments are conducted on NVIDIA A40 and RTX 3090 environments. If the work is accepted, the code will be released in subsequent versions.

REFERENCES

Mehryar Abbasi, Hadi Hadizadeh, and Parvaneh Saeedi. Reinforcement learning for unsupervised video summarization with reward generator training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. Demonstrating specification gaming in reasoning models. *arXiv preprint arXiv:2502.13295*, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

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.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

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. In *International Conference on Learning Representations*, volume 2026, pp. 132869–132895, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Yuming Li, Yikai Wang, Zhongyu Zhao, Ming Lu, Qi She, Shanghang Zhang, et al. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 20839–20866, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

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

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. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

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. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026b.

Ming Liu, Yunbei Zhang, Shilong Liu, Liwen Wang, and Wensheng Zhang. Wan-r1: Verifiable-reinforcement learning for video reasoning. *arXiv preprint arXiv:2603.27866*, 2026c.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37:134387–134429, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

Yash Savani, Branislav Kveton, Yuchen Liu, Yilin Wang, Jing Shi, Subhojyoti Mukherjee, Nikos Vlassis, and Krishna Kumar Singh. Stepwise credit assignment for grpo on flow-matching models. *arXiv preprint arXiv:2603.28718*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 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.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort. In *International Conference on Learning Representations*, volume 2026, pp. 96956–96981, 2026a.

Yifan Wang, Yanyu Li, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-drf: Differentiable reward flow for video diffusion fine-tuning. *arXiv e-prints*, pp. arXiv–2601, 2026b.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *Advances in Neural Information Processing Systems*, 38:171632–171668, 2026.

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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Haotian Ye, Kaiwen Zheng, Jiashu Xu, Puheng Li, Huayu Chen, Jiaqi Han, Sheng Liu, Qinsheng Zhang, Hanzi Mao, Zekun Hao, et al. Data-regularized reinforcement learning for diffusion models at scale. *arXiv preprint arXiv:2512.04332*, 2025.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *Advances in Neural Information Processing Systems*, 38:122647–122676, 2026.

APPENDIX

A Algorithmic Details and Experimental Settings

Algorithm 1 Training Procedure of LA-Framework

Require: Policy π_θ; length pool $\mathcal{L}_{\text{pool}} = \{L_k\}_{k=1}^{K}$; prompt distribution $\mathcal{D}$; reward model $R(\cdot)$.
Require: Hyperparameters: group size G, timesteps $\mathcal{T}$ ($|\mathcal{T}| = T$), branch points M, branch rollouts B, prior temperature β, clip parameter ϵ.
Ensure: Optimized policy π_θ.

// Stage 0: Offline Calibration of Credit Weights (re-executed every R steps if enabled)
Sample a small calibration set of prompts spanning $\mathcal{L}_{\text{pool}}$ and roll out trajectories with the initial LoRA policy π_θ
for each length $L_k \in \mathcal{L}_{\text{pool}}$ **do**
  // Step 1: Prior-guided Allocation
  Evaluate length prior $w_{\text{prior}}(t, L_k; \beta)$ over $\mathcal{T}$ (Eq. equation 2)
  // Step 2: Sensitivity-driven Refinement
  Select M branch timesteps $\{t_m\}_{m=1}^{M} \subset \mathcal{T}$ uniformly
  for each calibration trajectory **and** branch point $t_m \in \{t_m\}_{m=1}^{M}$ **do**
    Roll out B completions $\{x_0^{(b)}\}_{b=1}^{B}$ from the shared prefix x_{t_m} via diffusion
    Evaluate branch rewards $\{R_{t_m,L_k}^{(b)}\}_{b=1}^{B}$ and compute sensitivity $S(t_m, L_k)$
  end for
  Interpolate $S(t, L_k)$ across all $t \in \mathcal{T}$, normalize, and fuse with w_{prior} to obtain and freeze $W(t, L_k)$ (Eq. equation 4)
end for
for each training iteration **do**
  Sample a batch of prompts $\mathcal{B} = \{c_j\}_{j=1}^{|\mathcal{B}|} \sim \mathcal{D}$
  for each prompt $c_j \in \mathcal{B}$ **and** length $L_k \in \mathcal{L}_{\text{pool}}$ **do**
    // Phase 1: Multi-length Rollout & Advantage Computation
    Sample G candidates $\{x_0^{(j,k,i)}\}_{i=1}^{G}$ by denoising $x_1 \sim \mathcal{N}(0, I)$ via $\pi_\theta(\cdot \mid L_k)$ over $\mathcal{T}$
    Compute rewards $r_{j,k,i} = R(x_0^{(j,k,i)}, c_j)$
    Compute group-relative advantage $A_{j,k,i}$ within the same-prompt, same-length group
    // Phase 2: Apply Precomputed Length-conditioned Weights
    Compute stepwise advantages $\tilde{A}_{j,k,i,t} = W(t, L_k) \cdot A_{j,k,i}$ for all $t \in \mathcal{T}$ using the latest frozen weights
  end for
  // Phase 3: Policy Optimization
  Update policy parameters θ via gradient ascent on the clipped objective $\mathcal{L}_{\text{policy}}(\theta)$
end for
return π_θ

(1) Full Training Procedure of LA-Framework: Algorithm 1 summarizes the complete training procedure of LA-Framework, using the same notation as Sec. 3. The procedure explicitly decouples credit assignment from policy optimization. In *Stage 0*, the length-conditioned credit weights are calibrated once before training: sparse branching generates B continuations at M uniformly selected timesteps over a small set of calibration trajectories, all sharing the complete prefix; the resulting branch reward deviation estimates the length-level sensitivity $S(t, L)$, which is fused with the length-conditioned prior to produce the timestep weights $W(t, L)$, frozen thereafter. Each training iteration then consists of two phases. In the *multi-length rollout phase*, the policy generates G candidates for every prompt–length pair (c_j, L_k), and advantages are normalized strictly within same-prompt, same-length groups (Eq. equation 6), preventing different lengths from competing over absolute reward magnitudes. In the *policy optimization phase*, the precomputed weights modulate the group advantages into the stepwise advantage $\tilde{A}_{j,k,i,t} = W(t, L_k) \cdot A_{j,k,i}$ for the clipped update (Eq. equation 7), without any additional rollout or scoring.

Table 2: Mapping from evaluation dimensions to underlying metrics.

Dimension	Underlying metrics
TA Text–video alignment	Overall consistency (VBench); Text–video alignment (VideoScore)
VQ Visual quality	Aesthetic quality (VBench); Visual quality (VideoScore)
MQ Motion quality	Subject consistency (VBench); Motion smoothness (VBench); Temporal consistency (VideoScore); Dynamic degree (VideoScore)
FC Factual consistency	Factual consistency (VideoScore)

Since branching is confined to Stage 0, its cost—$B\sum_{m=1}^{M}(T-t_m)$ additional denoising steps per calibration trajectory, roughly $BMT/2$ with uniformly spaced branch points—is incurred only once before training and is independent of both the group size G and the length of the training run; no branching is performed during training, and the per-step training cost is the same as vanilla GRPO run at each length separately, with the frozen weights applied as a closed-form lookup. The calibration cost also remains far below full branching, which would require $T \times B$ branch continuations instead of $M \times B$. Since all branches share the complete prefix, no preceding frames are regenerated, and the extra memory footprint is limited to decoding the $M \times B$ completed branch videos for reward computation. When periodic recalibration is enabled, the calibration cost C_{cal} recurs only every R steps, so its amortized per-step overhead C_{cal}/R decreases linearly with the interval and is negligible for moderate R.

(2) Detailed Experimental Settings: Training details. All of our experiments are conducted on NVIDIA A40 and RTX 3090 GPUs. Training is performed in mixed precision with gradient checkpointing, allowing the full multi-length pipeline to fit within a single device. We adopt LoRA (rank 8) for parameter-efficient training, and the offline calibration of Stage 0 is executed once with the initial LoRA policy before training. Unless otherwise specified, we use group size G=8, M=4 branch points with B=4 branch rollouts. We optimize with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-4}) and the prior temperature is set to β=0.5, selected on validation rewards. Following Sec. 3.3, videos of all lengths share the same T=20 generation timesteps. For the main cross-length experiments, the length pool is set to $\mathcal{L}_{pool} = \{29, 49, 81\}$ frames, deliberately disjoint from the evaluation lengths $\{33, 57\}$, which are held out to assess cross-length generalization.

Comparison protocol. For the comparisons in Sec. 4.4, each baseline is trained with its official recipe at 49 frames, following the conventional protocol where training and evaluation share the same length. LA-Framework is instead trained once on the compact pool $\mathcal{L}_{pool} = \{25, 33, 49\}$ frames, whose maximum duration matches the baselines' training regime. All methods are evaluated at both 29 and 49 frames, where the 29-frame length is unseen for every method, though it lies within the span of our pool. Table 1 reports the average over the two lengths, and the 49-frame in-distribution results alone are reported in Appendix E.

B Extended Analysis of Length-Scale Inconsistency

The main text (Sec. 4.2) establishes the existence of Length-Scale Inconsistency at the dimension level. This section extends the analysis in three aspects: (1) the dimension-level aggregated effects across all five proxy rewards (Fig. 5); (2) per-metric effects with FDR-controlled significance, which localize the divergence to individual evaluation metrics (Fig. 6); and (3) a complete case study of the Aesthetic reward (Fig. 7). Table 2 summarizes the mapping between the four evaluation dimensions and the nine underlying metrics from VBench and VideoScore. All experiments follow the setup detailed in Sec. 4.1.

The inconsistency is systematic across proxy rewards. As shown in Fig. 5, the divergence between short- and long-video training is consistently observed under all five proxy rewards: short-video training produces predominantly positive effects on text–video alignment and visual quality, while long-video training causes broad degradations on visual quality, motion quality, and factual consistency. The asymmetry in magnitude is also evident: short-video training yields mild deviations (below 0.3 SD), whereas long-video training frequently exceeds -0.3 SD, indicating that long-video RL suffers substantially more severe hacking than its short-video counterpart.

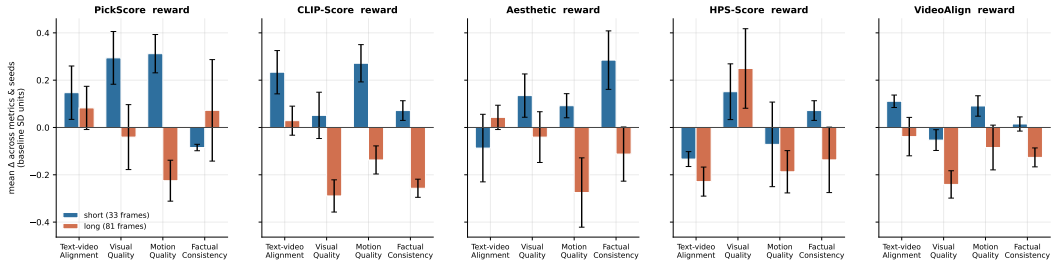


Figure 5: **Dimension-level effects of fixed-length RL across five proxy rewards.** For each proxy reward, bars report the mean change (after − before, in baseline SD units, averaged over underlying metrics and random seeds) of the four evaluation dimensions when training on short (33 frames) versus long (81 frames) videos. Error bars denote the standard error across metrics and seeds. Short-video training yields mild positive effects, whereas long-video training causes severe and broad degradations.

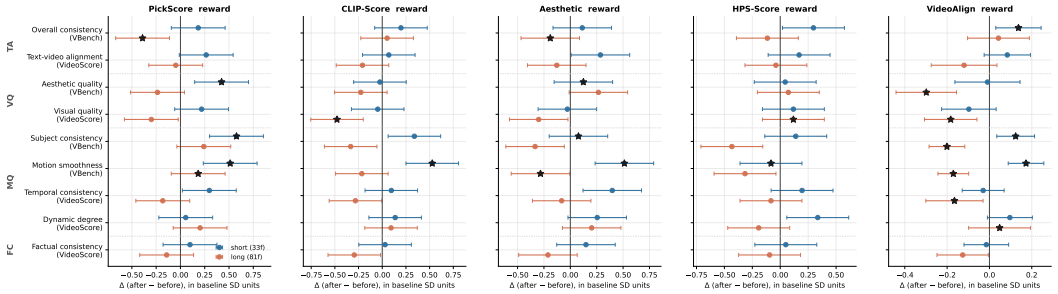


Figure 6: **Per-metric effects of fixed-length RL across five proxy rewards.** Each row corresponds to one evaluation metric, grouped by dimension (TA, VQ, MQ, FC); dots and error bars denote the mean change (after − before, in baseline SD units) with 95% confidence intervals for short (33 frames) versus long (81 frames) training. Black stars mark FDR-significant effects ($q < 0.05$). Significant short-training gains concentrate on frame-level metrics, whereas significant long-training degradations concentrate on cross-frame consistency metrics.

The hacking direction follows the bias of each reward. Beyond the canonical pattern, the aggregated view reveals reward-dependent deviations: under the HPS-Score reward, short-video training even degrades text–video alignment, whereas under the Aesthetic reward it improves factual consistency. Such deviations indicate that fixed-length RL exploits reward-specific blind spots rather than universally improving generation quality, consistent with the reward-hacking nature of the phenomenon.

Divergence localizes to frame-level versus cross-frame metrics. The per-metric analysis in Fig. 6 further reveals a systematic structural pattern behind the dimension-level trend. Significant short-training gains concentrate on frame-level metrics that are attainable through static, per-frame appeal, such as aesthetic quality and motion smoothness, whereas significant long-training degradations concentrate on cross-frame consistency metrics, such as overall consistency, subject consistency, and temporal consistency. Notably, the same metric can move in opposite directions under the two lengths (e.g., motion smoothness under the Aesthetic reward), because a short clip can inflate smoothness by suppressing motion while a long video cannot avoid genuine motion. This provides direct evidence that fixed-length RL optimizes reward-correlated proxies rather than true generation quality.

Metric-level case study of the Aesthetic reward. Figure 7 presents the complete per-metric results for the Aesthetic reward. Two observations stand out. First, short-video training improves 8 of 9 metrics, dominated by motion smoothness ($+0.51$) and temporal consistency ($+0.40$), and the coherent movement of metrics from the two independent evaluators (VBench and VideoScore) supports the robustness of the measurement. Second, long-video training degrades 7 of 9 metrics, while the only gains are confined to aesthetic quality ($+0.27$) and dynamic degree ($+0.20$), precisely the attributes favored by an appearance-oriented reward model. This coexistence of reward-aligned gains

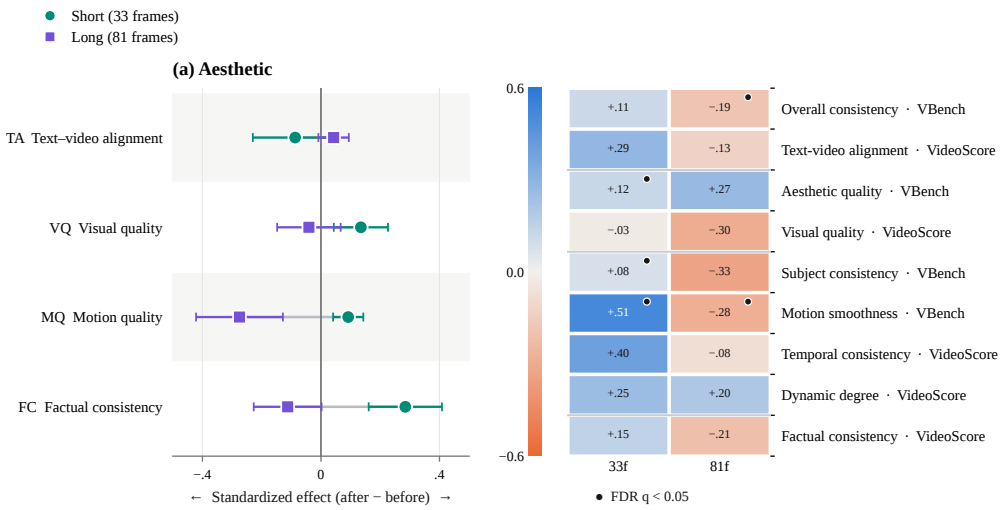


Figure 7: **Case study of the Aesthetic reward: per-metric effects at both training lengths.** Left: standardized effects (after − before) of the four evaluation dimensions under short (33 frames) versus long (81 frames) training, with 95% confidence intervals. Right: signed effects of all nine underlying metrics; blue and red cells indicate improvements and degradations, and dots mark FDR-significant effects ($q < 0.05$). Short-video training improves 8 of 9 metrics, whereas long-video training degrades 7 of 9, with the only gains confined to aesthetic quality ($+0.27$) and dynamic degree ($+0.20$).

with collateral degradation on consistency (e.g., subject consistency -0.33 and motion smoothness -0.28) is the signature of reward hacking rather than genuine quality improvement.

Overall, the extended analyses show that Length-Scale Inconsistency is systematic across proxy rewards, structurally localized to the frame-level versus cross-frame distinction, and characterized by reward-aligned inflation accompanied by collateral degradation, which reinforces the main-text conclusion that fixed-length RL training is fundamentally misaligned across video.

C Extended Results of Unified Joint-Length Training

The main text (Sec. 4.3) quantifies the cross-length consistency of LA-Framework via deviation and correlation statistics. This section presents the complete underlying results for the four proxy rewards (PickScore, CLIP-Score, VideoAlign, and HPS-Score). Figure 8 reports the dimension-level aggregated effects, and Figure 9 further resolves them into per-metric effects. For the unified model, effects are measured at all three evaluation lengths (33, 57, and 81 frames) of a single trained policy. All experiments follow the setup in Sec. 4.1, and metrics follow the mapping in Table 2.

Effects are consistent and predominantly positive across evaluation lengths. As shown in Fig. 8, 37 of the 48 reward–dimension effects are positive, and the cross-length spread remains below 0.10 SD for 14 of the 16 reward–dimension pairs. The few negative entries are all small ($|\Delta| \leq 0.17$) and uniform in sign across the three evaluation lengths, indicating that they reflect the intrinsic bias of the corresponding reward model rather than length-specific hacking.

Consistency metrics recover from long-video hacking. The per-metric analysis in Fig. 9 shows that the cross-frame consistency metrics most severely damaged by long-video fixed-length training (Sec. B)—such as subject consistency and temporal consistency—now remain positive or stable. Representative examples include subject consistency under the VideoAlign reward ($+0.19$, $+0.32$, $+0.33$) and motion smoothness under the HPS-Score reward ($+0.07$, $+0.10$, $+0.16$), whose effects are not only positive but grow mildly with video length, matching the increasing demand for temporal coherence of longer videos. Under the PickScore reward, overall consistency ($+0.34$/$+0.31$) and text–video alignment ($+0.30$/$+0.27$) likewise hold stable across lengths.

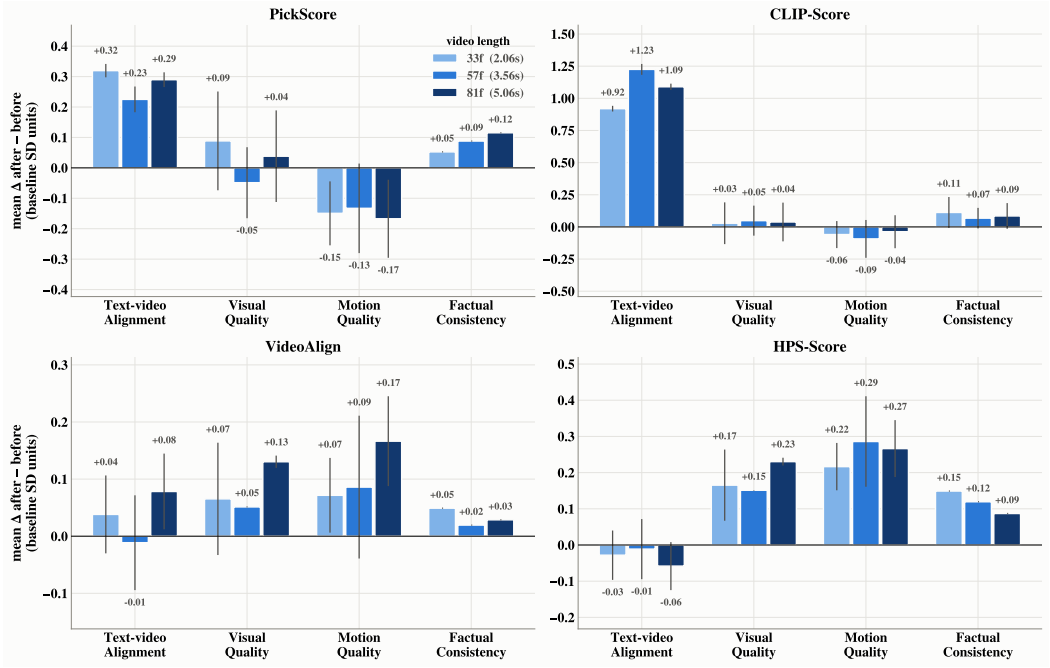


Figure 8: **Dimension-level effects of unified joint-length training across four proxy rewards.** For each proxy reward, bars report the mean change (after − before, in baseline SD units, averaged over underlying metrics and random seeds) of the four evaluation dimensions, evaluated at 33 (2.06 s), 57 (3.56 s), and 81 frames (5.06 s) of the unified model. Error bars denote the standard error across metrics and seeds. For nearly all reward-dimension pairs, the effect profile is nearly identical across the three evaluation lengths and predominantly positive, in sharp contrast to the length-divergent and frequently negative effects of fixed-length training.

Overall, these extended results confirm at both the dimension and metric levels that LA-Framework replaces the length-divergent optimization of fixed-length RL with a shared set of positive, length-stable effects, providing the statistical basis for the consistency analysis in the main text.

D Ablation Studies and Efficiency Analysis.

(1) Component-wise Ablation. To verify the necessity of each core design in LA-Framework, we conduct ablation experiments on Wan2.1-1.3B with the length pool $\{29, 49, 81\}$ frames, using VideoAlign as the reward model. We evaluate three variants by individually removing one of the three key components: (1) **w/o Sparse Branching**, which sets the empirical sensitivity to a constant ($\widehat{S}(t, L) \equiv 1$); by Eq. equation 4, the timestep weights then reduce to the length-conditioned prior alone, $W(t, L) = w_{\text{prior}}(t, L)$, so that credit assignment relies purely on the hand-crafted prior without any reward-feedback refinement; (2) **w/o Length-conditioned Prior**, which removes $w_{\text{prior}}(t, L)$ in Eq. equation 2 and relies solely on the empirical sensitivity $\widehat{S}_i(t, L)$ to form the timestep weight; and (3) **w/o Intra-group Normalization**, which replaces the length-conditioned intra-group normalization in Eq. equation 6 with a standard normalization computed jointly over the entire joint-length batch. Table 3 reports the average VBench and VideoScore metrics.

Removing any of the three components leads to a consistent degradation in both generation quality and cross-length consistency, confirming that each design contributes complementary and non-redundant benefits. Specifically, **w/o Sparse Branching** causes the largest drop in absolute quality (e.g., -2.61 on Subject Consistency and -2.83 on VQ). Note that this variant retains the length-conditioned prior ($W = w_{\text{prior}}$); the drop therefore isolates the value of the sensitivity refinement: the prior encodes only a coarse, reward-agnostic allocation over timesteps, and without branch-based refinement the gradient budget remains misaligned with the timesteps that actually drive reward changes, the credit misalignment discussed in Sec. 3. The fully uniform variant ($W(t, L) \equiv 1$,

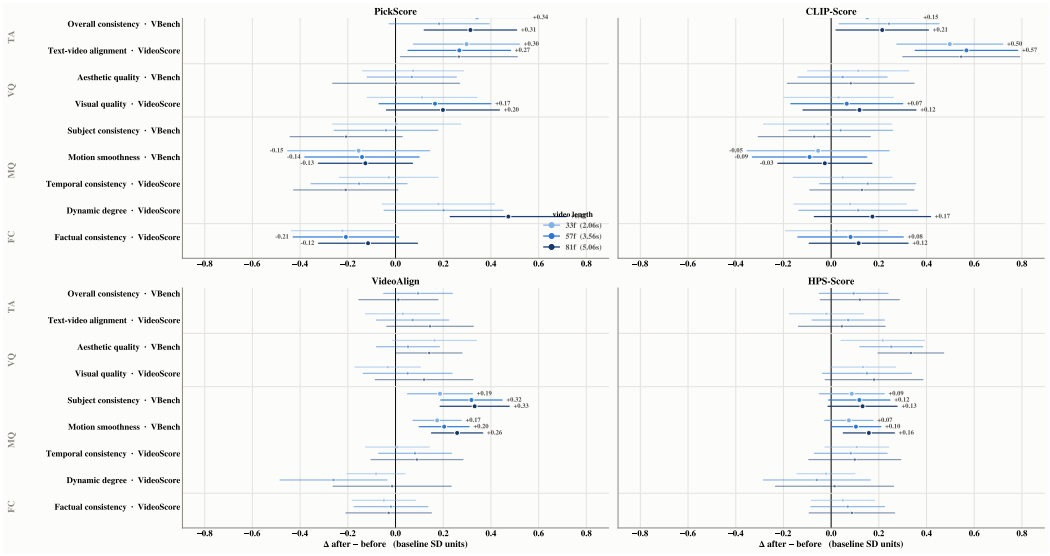


Figure 9: **Per-metric effects of the unified model across four proxy rewards.** Each row corresponds to one evaluation metric, grouped by dimension (TA, VQ, MQ, FC); dots and intervals denote the mean change (after − before, in baseline SD units) with confidence intervals at the 33-, 57-, and 81-frame evaluation lengths (light to dark blue). Annotated values mark selected effects. The sign of each metric-level effect is stable across evaluation lengths, and the consistency metrics that collapsed under long-video fixed-length training now remain positive or stable.

Table 3: Ablation study on Wan2.1. "Cross-Length Spread" denotes the mean standardized-effect spread across lengths; all other columns are absolute scores (higher is better). **Bold** indicates the best result.

Method	VBench		VideoScore		Cross-Length
	Subject Cons.	Motion Smooth.	VQ	TC	Spread↓
LA-Framework (full)	**90.92**	**94.95**	**77.85**	**80.82**	**0.087**
w/o Sparse Branching	87.31	92.47	75.02	78.14	0.156
w/o Length Prior	88.68	93.88	76.41	79.35	0.121
w/o Intra-group Norm.	84.94	93.02	74.88	77.69	0.183

removing both components) degenerates to standard GRPO-style advantage broadcasting, whose cross-length behavior is already characterized by the fixed-length baselines in Sec. 4.3. **w/o Length Prior** shows a smaller but non-trivial degradation, indicating that while empirical branch sensitivity alone captures part of the length-dependent structure, it remains noisy, as it is estimated from only $M \times B$ sparse branch rollouts. The prior provides a useful regularization that stabilizes the weight tree, particularly for the longest length in the pool. **w/o Intra-group Normalization** yields the largest increase in cross-length spread (0.183 vs. 0.087), confirming that normalizing advantages across the full joint-length batch (rather than within each length group) allows length-specific reward magnitude differences to be misread as quality differences, reintroducing the optimization conflicts that the length-conditioned grouping is designed to eliminate. Together, these results verify that sparse branching, the length-conditioned prior, and intra-group normalization each play a distinct and necessary role in achieving both high generation quality and consistent optimization across length scales.

(2) Efficiency Analysis. A natural concern is that the additional branching rollouts required to estimate $S(t, L)$ may introduce substantial training overhead compared to vanilla GRPO. To isolate this cost, we compare LA-Framework against vanilla GRPO under an identical setup (Wan2.1-1.3B, $G = 8$ candidates per prompt, single length $L = 49$ frames), varying the number of branch points M and branches per point B; the multi-length joint training itself scales the rollout and reward cost

Table 4: Training time and GPU memory of LA-Framework relative to vanilla GRPO, measured in the *online* variant where the sensitivity estimation is re-executed at every training step (an upper bound; our default offline calibration performs this estimation once before training). "Overhead" is the relative increase over vanilla GRPO (memory overhead in parentheses). Sparse branching with shared prefixes keeps the extra rollout cost small, since only the post-branch continuation ($M \times B$ short segments) is regenerated instead of the full trajectory.

Method	M	B	Time/Step (s)	Overhead	Peak Mem. (GB)
Vanilla GRPO	–	–	18.3	–	18.4
LA-Framework (M=2)	2	4	19.7	+7.7%	19.1 (+3.8%)
LA-Framework (M=4)	4	4	21.2	+15.8%	19.6 (+6.5%)
LA-Framework (M=4)	4	8	23.6	+29.0%	20.6 (+11.9%)

by the pool size K, the same factor as applying vanilla GRPO at each length separately. Results are reported in Table 4.

As shown in Table 4, even in the online variant, the default configuration ($M = 4$, $B = 4$) increases the per-step training time by only 15.8% and the peak GPU memory by 6.5% relative to vanilla GRPO; the length prior (Eq. equation 2) is a closed-form normalization with negligible cost. In practice, the sensitivity estimation is orchestrated as a standalone offline calibration (refreshed only every R steps when enabled): using the initial LoRA policy, we run sparse branching once on a small calibration set of $\sim$100 prompts spanning the length pool, derive $W(t, L)$, and freeze them. The subsequent large-scale training directly applies these precomputed weights at every step, in the default configuration without any additional branch rollouts, VAE decoding, or reward scoring, so the rollout cost analyzed above is paid only once and amortized over the entire run, and the per-step training overhead of LA-Framework is negligible. Even under the most aggressive online setting we tested ($M = 4$, $B = 8$), the overhead remains below 30% in time and under 12% in memory, well within the acceptable range given the substantial gains reported in Sec. 4.3 and Table 3. At inference time, LA-Framework introduces no additional cost, as the credit weights are discarded after training and the model performs standard single-pass denoising. Empirically, re-running the calibration at the midpoint of training yields sensitivity profiles highly correlated with the initial ones (Pearson $\rho = 0.90$), indicating that the frozen weights remain representative throughout training.

E Additional Comparison Results and Visualizations

(1) Results under the single-length evaluation. As a complementary view to the averaged results in Table 1, Table 5 reports the evaluation restricted to 49-frame videos, i.e., the conventional single-length protocol where all methods generate 49-frame videos at test time.

Reinforcement learning improves most metrics over the base models for every method in Table 5, confirming the general effectiveness of RL for video generation. The comparison is therefore not about whether RL helps, but about which method converts the reward signal into genuine quality most faithfully. In this regard, the baselines are uneven: their gains concentrate on the VideoScore dimensions closely related to the VideoAlign reward itself, while the independent VBench evaluation exposes collateral degradations (e.g., DanceGRPO reduces VBench imaging quality on Wan2.1, 44.29 vs. 45.64 of the base model), reproducing the reward hacking signature identified in Sec. 4.2.

Even under this single-length protocol, LA-Framework ranks first on five of nine metrics on both base models and never falls below the base model on any dimension. Notably, its lead is visibly narrower here than under the two-length average (five of nine versus 17 of 18 first places), which directly supports the protocol argument in Sec. 4.4: the length-specific gains of fixed-length baselines appear strong at a single favored length but are offset at the other, whereas the gains of LA-Framework are length-stable and thus accumulate under the averaged evaluation.

(2) Qualitative comparison. Figures 10 and 11 present qualitative results under the same 49-frame setting as Table 5, with several evenly sampled frames shown for each generated video. The observed patterns correspond closely to the quantitative findings. The baselines exhibit the two artifact families identified in Sec. 4.2: DanceGRPO over-saturates colors and over-sharpens tex-

Table 5: Comparison under the conventional single-length evaluation (all methods generate 49-frame videos; all trained with the VideoAlign reward). Best in **bold**, second best underlined. LA-Framework ranks first on five of nine metrics on both base models without sacrificing any dimension, and its advantage widens further under the two-length averaged evaluation in Table 1.

Model	Method	VBench ↑				VideoScore ↑				
		Subject consistency	Aesthetic quality	Motion smoothness	Imaging Quality	VQ	TC	DD	TVA	FC
Wan2.1	Baseline	87.15	46.22	91.79	45.64	70.49	77.71	66.38	68.70	58.61
	DanceGRPO	**91.67**	47.66	94.28	44.29	74.23	78.25	75.04	69.17	64.43
	FlowGRPO	90.09	45.78	95.17	45.23	73.69	80.10	73.53	**70.93**	67.25
	DDRL	89.78	47.04	**95.32**	45.94	75.92	**81.98**	72.65	68.85	68.50
	Ours	90.92	**48.33**	94.95	**46.07**	**77.85**	80.82	**76.76**	70.29	**69.69**
CogVideoX	Baseline	84.76	42.57	88.07	38.33	64.21	73.99	63.02	62.68	52.19
	DanceGRPO	86.05	46.20	89.29	41.11	66.76	75.12	67.10	68.46	59.62
	FlowGRPO	88.43	45.01	90.45	40.44	**68.65**	77.27	68.61	67.33	58.55
	DDRL	90.65	**46.94**	90.67	41.02	66.07	**78.00**	67.71	**69.49**	60.65
	Ours	**91.29**	46.73	**91.44**	**42.81**	67.93	77.51	**69.90**	68.51	**60.77**

tures (static exploitation), FlowGRPO introduces striped color artifacts, and DDRL produces severe motion blur, deformed subjects, and frames where the subject nearly disappears (motion artifacts). In contrast, LA-Framework maintains natural colors, stable subject identities, and coherent motion across frames, visually confirming that its gains in Table 5 reflect genuine quality improvements rather than reward-specific artifacts.

F Discussion

We discuss the main methodological concerns regarding attribution, compute alignment, the interpretation of the credit assignment, and the scope of our claims, and clarify what the current evidence does and does not establish.

The most critical concern is whether the improvements in Table 1 stem from the length-aware credit assignment or from the data diversity of multi-length training. Although a direct multi-length GRPO baseline is not included, three existing results partially separate the two factors. First, the ablation without intra-group normalization retains the full multi-length pipeline and the length-conditioned weights and changes only the normalization scope, yet its cross-length spread rises from 0.087 to 0.183 with consistent quality drops, showing that length-consistent optimization does not follow automatically from multi-length batching. Second, the single-length evaluation in Appendix E is the setting most favorable to the baselines: all of them are evaluated at their common training length of 49 frames, so no length shift is involved for them, whereas the averaged protocol additionally includes 29 frames, which is unseen for every method. Even under this favorable setting our method ranks first on five of nine metrics on both backbones and never falls below the base model, so an algorithmic advantage remains when the exposure asymmetry is absent; the margin then widens under the averaged protocol because the baselines' length-specific gains fail to transfer to 29 frames, which is the length-divergent behavior established in Sec. 4.2 rather than a protocol artifact. Third, the artifact families of fixed-length RL are absent in the qualitative results of the unified model. A direct multi-length GRPO comparison would further isolate the two factors and is left as future work; the current evidence does not by itself establish a unique causal attribution, but it supports our reading of multi-length data as the substrate that exposes the problem and of the length-aware credit assignment as the mechanism that converts it into consistent gains.

A related concern is compute alignment. Joint training samples $|\mathcal{L}_{pool}| \times G$ candidates per prompt, so its per-iteration sampling and reward cost exceeds that of single-length baselines, and the main comparison is step-matched rather than FLOPs-matched. The main table should therefore be interpreted as a comparison of optimization procedures under their stated protocols rather than as a compute-normalized one. We note two mitigating points: multi-length sampling is inherent to the

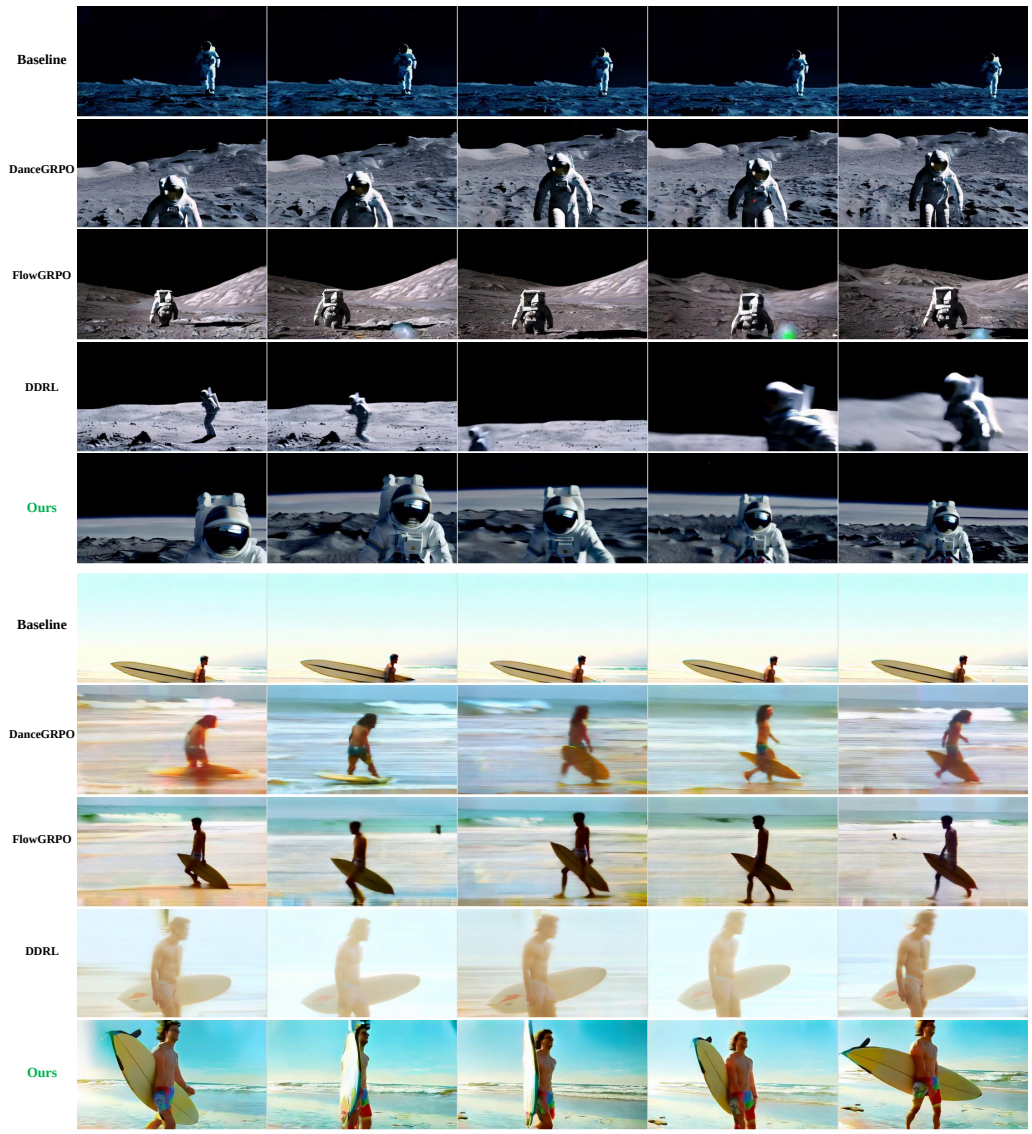


Figure 10: Visual comparison under the same 49-frame setting as Table 5 (VideoAlign reward). Each row denotes one method, and each column shows an evenly sampled frame of the generated video.

problem definition, since any method that optimizes across lengths must consume multi-length data and the baselines could equally be run on the same pool; and the extra samples are shared across lengths of the same prompt, so they mainly extend length coverage rather than prompt-level optimization depth, although they do incur real computation and may contribute part of the observed gains. A compute-matched protocol, for instance by adjusting baseline steps until total generated frames align, would be the appropriate test and remains a limitation of the current comparison. The consistency analysis and the comparison adopt different pools by design: the former uses a wider pool with held-out lengths to study length-scale behavior, while the latter uses a compact pool matched to the baselines' regime.

On the interpretation of $S(t, L)$, we agree that the branch-based reward variance measures the responsiveness of future outcomes rather than the strict causal contribution of timestep t, since it also absorbs continuation stochasticity, downstream dynamics, and reward-model noise. Our claim is accordingly practical: $S(t, L)$ is an empirical and computationally cheap proxy for allocating the optimization budget, whose usefulness is established by ablations, and the current results do not

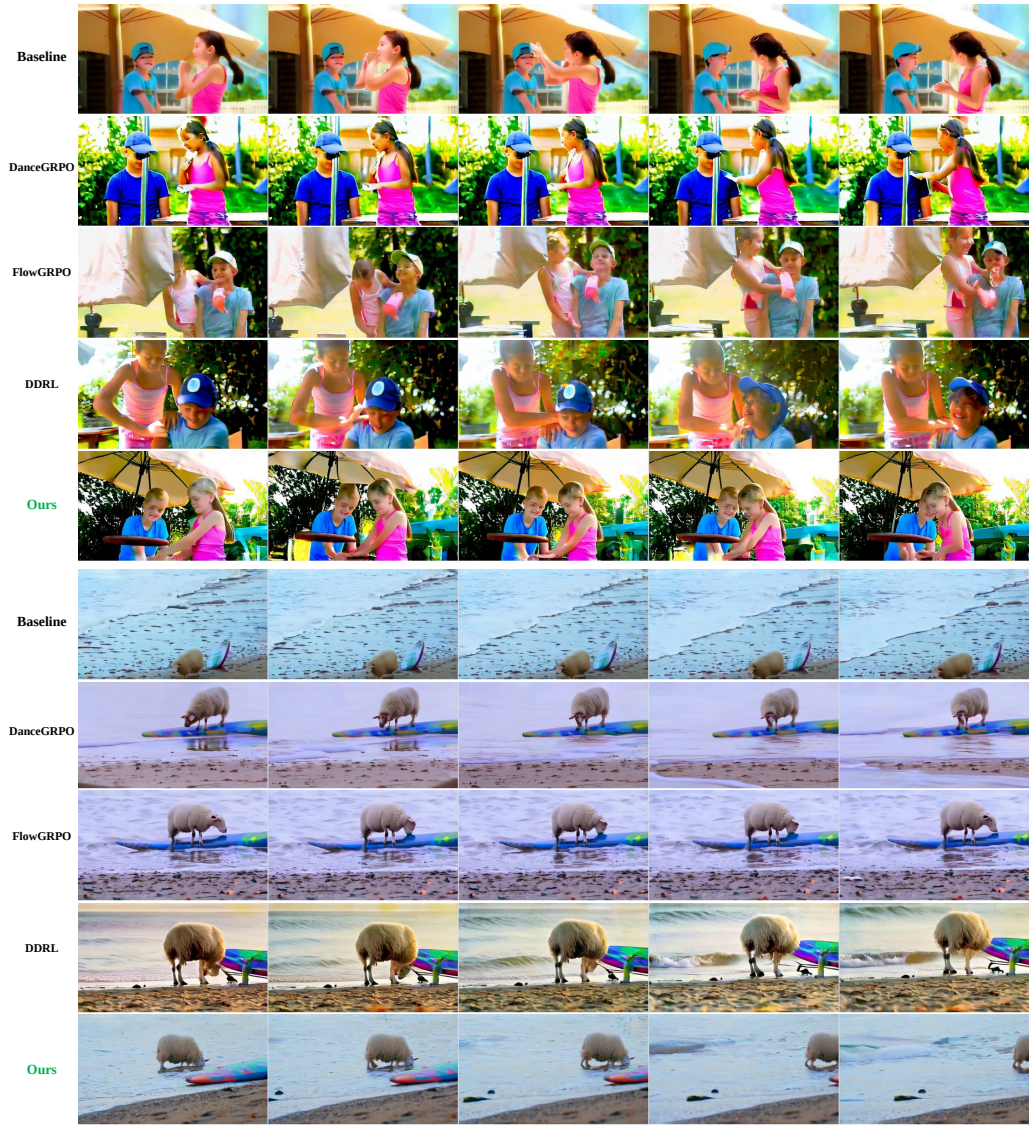


Figure 11: Additional qualitative comparison on more diverse prompts.

establish that it recovers the true causal credit of each timestep. Regarding the frozen weight tree, the midpoint recalibration correlation of ρ=0.90 indicates that the sensitivity profile is largely preserved as the policy drifts; the prior anchors the allocation to the coarse temporal ordering implied by video length, preventing noisy sparse estimates from dominating, and the normalized weights modulate relative budgets among timesteps rather than absolute gradient magnitudes. This evidence reduces, but does not eliminate, the risk of calibration staleness, and dynamic recalibration with a finite period is a natural extension.

The exponential prior is a design choice rather than a theoretical optimum. Its role is to provide a stable length-dependent initialization for noisy sparse sensitivity estimates, and the ablation supports this reading: removing the prior degrades performance moderately, whereas removing sparse branching or intra-group normalization causes larger drops. This establishes the usefulness of the prior as regularization but not the optimality or necessity of the exponential form, and a sensitivity analysis over β and alternative monotone parameterizations is left for future work.

Concerning the strength of the reward hacking claim, our most direct evidence establishes length-consistent optimization, while the suppression of length-specific hacking is an interpretation built

on three indirect observations: the divergent artifact patterns of fixed-length RL disappear under joint training, the mean effects become predominantly positive, and the residual deviations remain bounded. We therefore present hacking mitigation as a supported interpretation of these observations rather than a proven mechanism. We also acknowledge an alternative reading in which proxy rewards carry length-dependent biases that joint training averages out. Our framework is agnostic to this distinction: whether the inconsistency originates from optimization dynamics or from reward bias, reallocating the gradient budget in a length-aware manner reduces length-specific over-optimization, and both readings motivate the same design.

Regarding the scope of generalization, our results support cross-length generalization within and around the training range, verified on the held-out interpolated lengths; they do not establish extrapolation beyond the pool, and since the prior depends on $L/L_{\max}$, its behavior outside the pool is smooth by construction but untested. We therefore restrict the claim to diverse durations within and near the studied range and treat long-video extrapolation as future work.

On novelty, the key distinction from BranchGRPO, TempFlow-GRPO, and related stepwise credit assignment methods is that all lengths in our setting share the same number of generation timesteps ($T{=}20$), so the credit structure varies with the target video length itself rather than with the position along a fixed trajectory. Concretely, the same timestep may deserve a large budget for an 81-frame video, where it determines long-range motion, and a small budget for a 29-frame video, where local detail dominates; no prior stepwise method expresses this length-dependent credit, since they condition credit only on the timestep index within a single-length trajectory. The second contribution, multi-length joint optimization with within-length normalization, addresses the resulting interference across length groups, which does not arise in single-length training. This distinction defines the scope of our methodological contribution.

Finally, on evaluation and reporting, all quantitative results rely on automatic metrics. VBench is independent of the VideoAlign reward family and the qualitative results show no residual artifact patterns, but these checks do not replace human preference studies or training with an additional independent reward, so we limit the claim to improvements on the reported metrics. On efficiency, branching is executed once during calibration rather than at every training step, and the shared prefix reduces the additional continuations from $T \times B$ to $M \times B$ rollouts, so the reported online overhead is an upper bound rather than representative of the default schedule; the one-time calibration cost should nevertheless be counted when comparing total training compute.