FRAMEGRPO: INCENTIVIZING VIDEO REASONING VIA MULTI-FRAME ROLLOUTS AND TEMPORAL CREDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards, particularly Group Relative Policy Optimization (GRPO), has improved video reasoning in multimodal large language models. However, standard video GRPO varies only linguistic trajectories while conditioning every rollout in a group on the same sampled frames, which can produce homogeneous rewards and weaken group-relative learning when the shared visual evidence is insufficient or distracting. We propose **FrameGRPO**, a simple and effective reinforcement-learning framework that requires neither auxiliary networks nor changes to the model architecture. Specifically, FrameGRPO combines two lightweight components: (1) **Multi-Frame Rollouts**, which use only uniform temporal subsampling of a single maximal-frame decode to construct aligned views that diversify visual evidence and reasoning trajectories; and (2) **Gated Temporal Efficiency Credit**, which selectively reinforces a correct lower-ratio trajectory when all rollouts in a fixed full-ratio reference subset fail, without imposing an unconditional preference for fewer frames. Extensive experiments across diverse video understanding and reasoning benchmarks demonstrate robust overall improvements over standard GRPO, highlighting temporal-view diversification as an effective mechanism for exploration and credit assignment in video MLLMs.

1 INTRODUCTION

Recent progress in reinforcement learning with verifiable rewards (RLVR) has advanced model reasoning beyond the imitation of curated solution traces toward learning directly from outcome-level feedback (Shao et al., 2024; Guo et al., 2025). Group Relative Policy Optimization (GRPO) has become a particularly practical recipe because it learns from relative comparisons among responses to the same query without requiring a separate critic. This paradigm has rapidly expanded from language reasoning to multimodal and video models (Feng et al., 2026; Wang et al., 2026b; Fu et al., 2026). Recent video-reasoning systems further improve fine-grained visual–temporal credit assignment or allow models to acquire relevant evidence during inference (Wang et al., 2026d; Zhang et al., 2026a; Xie et al., 2025; Ding et al., 2026). Together, these advances establish RL post-training as a promising route from passive video recognition toward reasoning grounded in temporal evidence.

Despite this progress, standard video GRPO explores variation primarily in the language space while keeping the visual context fixed. Rollouts for the same video–question pair use the same frame sampling ratio and therefore the same temporal view, so stochastic decoding changes the linguistic trajectory but not the density of visual input. Yet the frame sampling ratio governs an important trade-off within the same decoded video: lower-ratio views contain fewer redundant or irrelevant frames and can reduce visual distraction, whereas higher-ratio views preserve more task-relevant temporal evidence. Fixing a single frame sampling ratio forces every rollout to inherit the same balance between evidence sufficiency and visual interference, making their outcomes more correlated. When a group becomes uniformly correct or incorrect, GRPO has little basis for relative comparison because all responses receive homogeneous rewards, causing the task-relevant learning signal to vanish. This *advantage collapse* wastes rollouts and can stall optimization (He et al., 2026a; Yu et al., 2026b; Zhang et al., 2026c). Existing difficulty-adaptive methods restore variation by modi-

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: **Training dynamics of FrameGRPO and GRPO.** The four panels track VSI-Bench accuracy, within-group reward standard deviation, and the prevalence of uniformly correct and uniformly incorrect rollout groups during optimization. Exposing each group to multiple frame sampling ratios yields more informative reward contrasts, mitigates group-level saturation at either outcome extreme, and translates into stronger evaluation performance. Shaded bands visualize the variability of each metric.

fying or regrouping questions (Gao et al., 2026), but they do not address the shared visual context that is distinctive to video reasoning.

We propose **FrameGRPO** to address this evidence-induced advantage collapse. Our key observation is that the frame sampling ratio is not merely a computational setting: for the same video and question, it changes which events are visible, how much distracting content is present, and which reasoning paths are elicited. FrameGRPO therefore turns the frame sampling ratio into a within-group exploration variable, allowing the policy to compare trajectories grounded in different but temporally aligned views. This controlled evidence diversity makes mixed outcomes more likely and preserves the relative signal that GRPO requires. Figure 1 illustrates this mechanism: compared with fixed-ratio GRPO, FrameGRPO maintains greater within-group reward variation, substantially reduces both all-correct and all-incorrect groups, and improves performance across video-reasoning benchmarks.

FrameGRPO consists of two complementary modules. The **Multi-Frame Rollouts** module first decodes each video once at the maximum sampled frame count, assigns the resulting full-ratio view to a fixed subset of rollout slots, and uniformly subsamples it for the remaining slots at randomly sampled frame sampling ratios. The same policy generates responses from these views, and a shared verifier evaluates them within one group, introducing diversity in both visual evidence and reasoning trajectories. **Gated Temporal Efficiency Credit** uses the outcomes of the fixed full-ratio subset to control credit assignment. If any full-ratio rollout is correct, it leaves the standard GRPO signal unchanged. Only when all full-ratio rollouts fail does it assign additional credit to successful lower-ratio trajectories, with greater emphasis on successes obtained at lower frame sampling ratios. Finally, every sampled completion is conditioned on the full-ratio input for policy and reference likelihood evaluation, transferring discoveries from complementary views into a single policy. Together, the two modules create useful cross-view comparisons without unconditionally favoring lower frame sampling ratios.

To validate the effectiveness of FrameGRPO, we evaluate it on seven diverse video understanding and reasoning benchmarks: Video-MME (Fu et al., 2025), TempCompass (Liu et al., 2024), VideoEval-Pro (Ma et al., 2025), MMVU (Zhao et al., 2025), VSI-Bench (Yang et al., 2025), NExT-GQA (Xiao et al., 2024), and Video-MMMU (Hu et al., 2026). The results show that FrameGRPO outperforms standard GRPO across all evaluated settings, demonstrating robust improvements across complementary aspects of video reasoning.

Our contributions are threefold:

- We propose **FrameGRPO**, a simple and effective reinforcement-learning framework that improves video reasoning.
- FrameGRPO explores multiple frame sampling ratios to diversify temporal evidence and introduces a cross-view credit assignment strategy that reinforces successful trajectories discovered from complementary ratio-conditioned views.

- Extensive experiments across seven video understanding and reasoning benchmarks demonstrate that FrameGRPO consistently surpasses standard GRPO, validating its effectiveness and robustness across diverse video tasks.

2 Related Work

2.1 MLLMs for Video Understanding

Video MLLMs must distill long, redundant streams into tractable evidence: sparse sampling can miss decisive events, whereas dense inputs increase cost and distraction (Zhu et al., 2026; Zou et al., 2026; He et al., 2026b). Existing work addresses this trade-off through compact visual memory (Chen et al., 2026a), retrieval (Luo et al., 2026; Zhang et al., 2026b; Bahrami et al., 2026), keyframe or adaptive sampling (Li et al., 2026b; Qin et al., 2026; Qiu et al., 2026), and token compression (Xu et al., 2026; Wu et al., 2026b). These strategies reduce redundancy before or during decoding but primarily optimize the evidence available to each individual reasoning trajectory. Complementary methods integrate evidence acquisition into reasoning through frame-wise chains of thought (Arnab et al., 2026), iterative search or temporal zooming (He et al., 2026b; Pan et al., 2026; Ding et al., 2026), and agentic tool use (Zeng et al., 2026; Xie et al., 2025; Zhang et al., 2026a; Jain et al., 2026). Temporal objectives, grounding post-training, and evidence-level rewards further improve temporal supervision (Wu et al., 2026c; Wang et al., 2026c; Li et al., 2026a; Meng et al., 2025; Cao et al., 2026). Unlike these approaches, FrameGRPO uses aligned multi-ratio views within each GRPO group as an informative signal for a single policy, requiring neither auxiliary modules nor additional frame-decoding computation.

2.2 Reinforcement Learning in MLLMs

RLVR methods such as GRPO optimize relative outcomes within groups without a learned critic (Shao et al., 2024; Guo et al., 2025; Yu et al., 2026b). In MLLMs, they support general or grounded visual reasoning (Yu et al., 2026a; Liu et al., 2026b; Sarch et al., 2026; Yao et al., 2026), active perception and tool use (Wu et al., 2026a; Chng et al., 2026; Zhu et al., 2025), and alternative optimization signals (Xiao & Gan, 2026; Yoon et al., 2026; Du et al., 2026). Video RL additionally requires deciding which temporal evidence to inspect. Prior systems establish reinforced fine-tuning (Feng et al., 2026; Wang et al., 2026a; Chen et al., 2026b), improve temporal grounding and sampling (Li et al., 2026c; Wang et al., 2026c; Qin et al., 2026), enable active navigation (Qiu et al., 2026; He et al., 2026b; Pan et al., 2026), or provide evidence-aware supervision (Wang et al., 2026d; Cao et al., 2026; Linghu et al., 2026). Nevertheless, when all rollouts share a fixed video view, insufficient or distracting evidence can still yield uniform rewards and erase the relative learning signal. Unlike prior remedies for uninformative uniform-reward groups (Yu et al., 2026b; Park et al., 2026; Gao et al., 2026; He et al., 2026a), FrameGRPO targets the video-specific cause—identical visual context across rollouts—using multiple frame-sampling views and Gated Temporal Efficiency Credit.

3 FrameGRPO

In this section, we introduce **FrameGRPO**, a simple and effective reinforcement-learning framework for improving video reasoning in multimodal large language models (MLLMs). To address the limited visual exploration of vanilla GRPO and the resulting risk of advantage collapse, FrameGRPO explores multiple frame sampling ratios and selectively reinforces successful trajectories from lower-ratio visual contexts via gated temporal efficiency credit when all rollouts in a fixed full-ratio reference subset fail. Section 3.1 reviews GRPO and its evidence-induced advantage collapse; Section 3.2 describes multi-frame rollout construction; and Section 3.3 introduces our gated credit mechanism. Figure 2 illustrates the overall pipeline.

3.1 Preliminaries: Group Relative Policy Optimization

Let V denote a sampled video view, q a question, and a its reference answer. Given the fixed multimodal prompt (V, q), vanilla video GRPO samples a group of G responses from the old policy, $o_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid V, q)$ for $i \in \{1, \ldots, G\}$. A shared verifier assigns each response a scalar reward $r_i =$

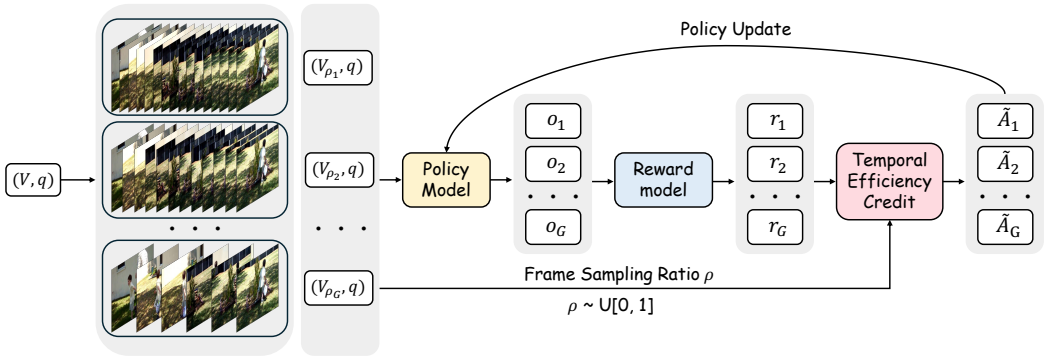


Figure 2: Overview of FrameGRPO. Given a video–question pair, we assign the full-ratio input to a fixed subset of rollout slots and independently sample $\rho_i \sim \mathcal{U}[0, 1]$ for the remaining slots to construct temporally aligned views V_{ρ_i}. A shared policy generates $\{o_i\}_{i=1}^G$, and a verifier assigns rewards $\{r_i\}_{i=1}^G$. The outcomes of the fixed full-ratio subset jointly determine the gate for Temporal Efficiency Credit, which produces the modified advantages $\{\widetilde{A}_i\}_{i=1}^G$. Policy optimization is conditioned only on the full-ratio input.

$R(o_i, a)$. In our instantiation, r_i is the sum of a correctness reward and a format reward; we retain the correctness signal separately for the gated credit introduced in Section 3.3. Let $\mathbf{r} = (r_1, \ldots, r_G)$ collect the rewards of all responses. GRPO avoids a learned critic and estimates the normalized advantage as

$$\widehat{A}_i = \frac{r_i - \mathrm{mean}(\mathbf{r})}{\mathrm{std}(\mathbf{r}) + \delta}, \tag{1}$$

where $\delta > 0$ is a small constant for numerical stability. The KL-regularized GRPO objective with a PPO-style clipped update is

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{\substack{(V,q,a)\sim\mathcal{D}, \\ \{o_i\}_{i=1}^G \sim \pi_{\theta_{\mathrm{old}}}(\cdot|V,q)}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\min \left\{ \frac{\pi_\theta(o_i \mid V, q)}{\pi_{\theta_{\mathrm{old}}}(o_i \mid V, q)} \widehat{A}_i, \right. \right. \right. \\ \left. \left. \left. \mathrm{clip}\left(\frac{\pi_\theta(o_i \mid V, q)}{\pi_{\theta_{\mathrm{old}}}(o_i \mid V, q)}, 1-\varepsilon, 1+\varepsilon \right) \widehat{A}_i \right\} - \beta \widehat{D}_{\mathrm{KL}}^{(i)}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}) \right) \right], \tag{2}$$

where π_θ is the current policy, $\varepsilon > 0$ sets the clipping range, π_{ref} is a frozen reference policy, and β controls the KL penalty.

Equation 1 relies on heterogeneous rewards. If every response is assigned the same reward, then both the centered rewards and their standard deviation vanish, yielding $\widehat{A}_i = 0$ and no reward-driven gradient. This failure mode is commonly referred to as *advantage collapse* (He et al., 2026a; Zhang et al., 2026c). It is particularly relevant to video GRPO: every rollout observes exactly the same frames, so stochastic decoding explores only the language trajectory while leaving the underlying visual evidence unchanged. An insufficient or distracting view can therefore make the entire group fail in a correlated manner. FrameGRPO addresses this modality-specific degeneracy by making the frame sampling ratio an additional within-group exploration variable.

3.2 Multi-Frame Rollouts for Temporal Exploration

Complementary temporal evidence. The frame sampling ratio controls which temporal evidence enters the model, not merely its computational cost. Because relevant and distracting events are unevenly distributed over time, changing this ratio can affect reasoning in a non-monotonic manner, as illustrated in Figure 3. We therefore vary the ratio across rollouts in each group and use the resulting complementary temporal views as a source of exploration.

Multi-frame rollout strategy. Let $f_\star$ denote the maximum number of sampled frames and V_ρ the view associated with a frame sampling ratio $\rho \in [0, 1]$; the full-ratio view V_1 therefore contains

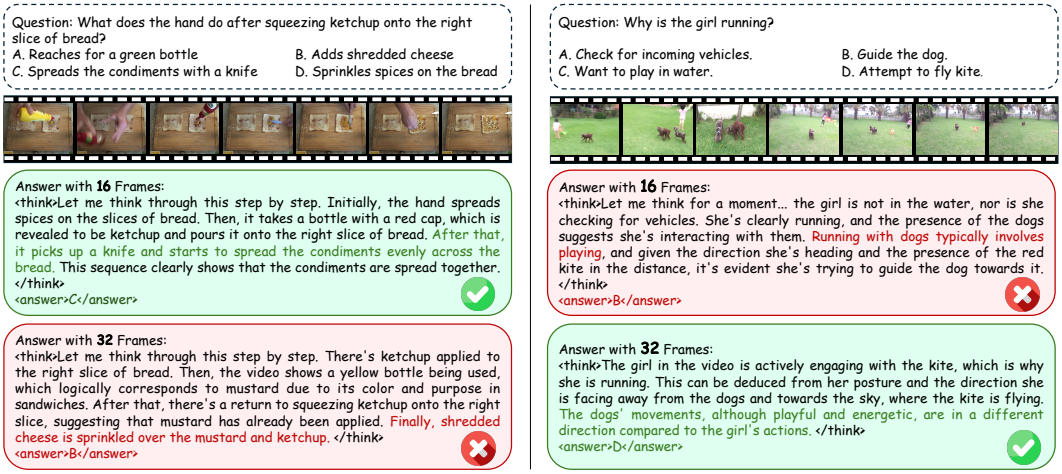


Figure 3: Complementarity across frame sampling ratios. *Left:* the lower-ratio 16-frame view captures the relevant condiment-spreading action, whereas the full-ratio 32-frame view is distracted by a subsequent event. *Right:* the higher-ratio view preserves the evidence required to associate the girl's motion with the kite, whereas the lower-ratio view misses this critical evidence.

$f_\star$ frames. To explore temporal evidence while retaining a reliable reference for credit assignment, every multi-frame group reserves a fixed subset of rollout slots for V_1. Let $\mathcal{I}_\star \subset \{1, \ldots, G\}$ index this subset and $K_\star = |\mathcal{I}_\star|$, where $1 < K_\star < G$. We set $\rho_i = 1$ for $i \in \mathcal{I}_\star$ and independently sample $\rho_i \sim \mathcal{U}[0, 1]$ for the remaining $G - K_\star$ slots. Each sampled ratio is mapped to an integer frame count $f_i = \lceil \rho_i f_\star \rceil$; after rounding, we use the realized ratio $\rho_i \leftarrow f_i / f_\star$ thereafter. We decode the video once at $f_\star$ frames and uniformly subsample f_i frames to construct V_{ρ_i}. All views cover the same temporal interval and preserve temporal order. The full-ratio slots share the complete sampled context V_1 but yield multiple stochastic rollouts, whereas the remaining slots provide ratio-conditioned exploration.

Conditioned on each view V_{ρ_i}, the old policy samples a rollout $o_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid V_{\rho_i}, q)$, which receives the verifier reward $r_i = R(o_i, a)$. Together, the full-ratio and exploratory subsets produce G responses $\{o_i\}_{i=1}^G$ grounded in complementary temporal contexts; the associated ratios $\{\rho_i\}_{i=1}^G$ and rewards $\{r_i\}_{i=1}^G$ are retained for subsequent credit assignment. In particular, the outcomes $\{(o_i, r_i)\}_{i \in \mathcal{I}_\star}$ provide the full-ratio reference required by Gated Temporal Efficiency Credit.

Ratio-conditioned views are used only during rollout collection; policy optimization is conditioned exclusively on the full-ratio input (V_1, q) to ensure stable learning and preserve the accuracy of the learned policy.

3.3 Gated Temporal Efficiency Credit

Multi-frame rollouts increase outcome diversity, but standard group normalization does not explicitly distinguish between a correct trajectory discovered from a lower-ratio view and those obtained from the fixed full-ratio reference subset. This subset makes the distinction observable in every group: when all of its rollouts fail but a lower-ratio view succeeds, the latter reveals a useful trajectory absent from the sampled full-ratio outcomes. We introduce **Gated Temporal Efficiency Credit** (TEC) to retain such discoveries without imposing an unconditional preference for lower frame sampling ratios.

We distinguish the scalar verifier reward r_i, which may combine correctness and format, from a binary correctness indicator $c_i = \mathbf{1}[o_i \text{ answers } q \text{ correctly}]$. The gate activates only when every rollout in the fixed full-ratio subset is incorrect:

$$g = \mathbf{1}\left[\max_{j \in \mathcal{I}_\star} c_j = 0\right]. \tag{3}$$

Table 1: **Performance Comparison on Video Understanding and Reasoning Benchmarks.** All results are reported as accuracy (%). The best result within each backbone is highlighted in bold.

Method	Video General Benchmark			Video Reasoning Benchmark			
	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
GPT-4o	71.9	75.1	34.2	75.4	34.0	-	61.2
Gemini 1.5 Pro	75.0	70.5	39.3	71.2	48.8	-	53.4
LLaMA-VID	-	45.6	-	-	-	-	-
LLaVA-OneVision-7B	58.2	69.5	-	49.2	32.4	-	33.8
Kangaroo-8B	56.0	62.5	-	-	-	-	-
VILA-1.5-8B	-	58.8	-	-	28.9	-	20.8
Video-R1-7B	60.0	73.2	42.6	63.4	36.7	77.3	51.8
VideoRFT	60.7	73.6	40.8	66.1	37.2	78.0	49.6
VideoChat-R1	59.6	73.8	39.3	64.3	36.3	78.3	51.3
Time-R1-7B	59.6	73.3	40.3	63.5	24.9	78.6	51.9
Main Results							
Qwen2.5-VL-7B-Instruct	57.4	72.9	38.6	61.4	31.9	76.9	51.7
+ SFT	53.8	69.5	36.5	61.6	31.9	70.0	43.6
+ GRPO	60.0	74.4	41.0	65.6	34.1	79.5	52.0
+ FrameGRPO	**61.3**	**74.9**	**42.4**	**66.7**	**37.9**	**79.6**	**54.2**
Qwen2-VL-7B-Instruct	57.4	70.1	42.0	59.2	23.2	77.2	31.3
+ SFT	52.3	67.4	39.1	56.8	**29.1**	67.7	35.8
+ GRPO	57.9	71.3	43.2	61.6	27.6	76.9	41.3
+ FrameGRPO	**58.4**	**72.3**	**44.8**	**62.2**	27.4	**77.3**	**43.6**

For each trajectory, TEC assigns the bonus

$$b_i = \alpha\, g\, c_i \log\left(\frac{1}{\rho_i}\right), \qquad \widetilde{A}_i = \widehat{A}_i + b_i, \tag{4}$$

where $\widehat{A}_i$ is the standard group-relative advantage in Equation 1 and $\alpha \geq 0$ controls the credit strength. We use $\alpha = 0.2$ by default. The inverse frame sampling ratio gives larger credit to a correct trajectory discovered from a lower-ratio view and is zero when $\rho_i = 1$.

We next define the complete FrameGRPO objective. Let $p_{\mathrm{MF}}(\boldsymbol{\rho})$ denote the group-level ratio assignment described above: $\rho_i = 1$ for $i \in \mathcal{I}_\star$, while the remaining ratios are independently sampled. The objective is

$$\mathcal{J}_{\mathrm{FrameGRPO}}(\theta) = \mathbb{E}_{\substack{(V,q,a)\sim\mathcal{D},\, \boldsymbol{\rho}\sim p_{\mathrm{MF}} \\ o_i\sim\pi_{\theta_{\mathrm{old}}}(\cdot|V_{\rho_i},q),\ i=1,\ldots,G}} \left[\frac{1}{G}\sum_{i=1}^{G} \left(\min\left\{ \frac{\pi_\theta(o_i \mid V_1, q)}{\pi_{\theta_{\mathrm{old}}}(o_i \mid V_1, q)}\widetilde{A}_i, \right. \right. \right. \\ \left. \left. \left. \mathrm{clip}\left(\frac{\pi_\theta(o_i \mid V_1, q)}{\pi_{\theta_{\mathrm{old}}}(o_i \mid V_1, q)}, 1-\varepsilon, 1+\varepsilon \right) \widetilde{A}_i \right\} - \beta \widehat{D}^{(i)}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}) \right) \right]. \tag{5}$$

The gate yields three desirable behaviors. If any rollout in the full-ratio reference subset succeeds, then $g = 0$ and $\widetilde{A}_i = \widehat{A}_i$, leaving the standard GRPO advantage unchanged. If all full-ratio reference rollouts fail but a lower-ratio view succeeds, only the correct lower-ratio trajectories receive additional credit. If every view fails, $c_i = 0$ for all trajectories and no bonus is introduced. Thus, TEC exploits a lower-ratio success only when it reveals a trajectory not discovered by any rollout in the fixed full-ratio subset; it does not reward a lower frame sampling ratio by default. Since the mechanism changes only the training-time advantages, it introduces no additional inference cost.

4 Experiments

4.1 Training Data

We construct our training set exclusively from the video portion of Video-R1-260K (Feng et al., 2026), which contains approximately 116K video samples collected from diverse open-domain sce-

narios. These data cover event progression, object motion, temporal dependencies, action understanding, and causal reasoning, providing broad supervision for spatiotemporal video understanding. Following prior work (Yu et al., 2026b; Zheng et al., 2026), we perform difficulty-based data curation by generating eight responses for each candidate question and estimating its difficulty from the empirical accuracy. We exclude questions with an accuracy of either zero or one, corresponding to cases where none or all of the responses are correct. After filtering, we obtain 8K examples as our final training set.

4.2 Implementation Details

We implement FrameGRPO on top of the Open-R1-Video (Wang et al., 2026b) reinforcement learning framework and conduct experiments with Qwen2.5-VL-7B and Qwen2-VL-7B as the base MLLMs, comparing standard GRPO against our proposed FrameGRPO for each backbone. For training efficiency, we cap the number of video frames at 32 and treat the uniformly sampled 32-frame sequence as the full-ratio view. Each FrameGRPO group assigns this fixed input to a subset $\mathcal{I}_\star$ of rollout slots; for every remaining slot, we independently draw $\rho_i \sim \mathcal{U}[0, 1]$ and uniformly subsample the full-ratio sequence to construct the corresponding temporally aligned view V_{ρ_i}. Each frame is processed with a maximum pixel budget of $256 \times 28 \times 28$. Each training run takes approximately 21 hours on a single node equipped with eight NVIDIA H20 GPUs.

4.3 Evaluation Benchmarks

We evaluate our models on seven benchmarks covering general video understanding and video reasoning. The video general benchmarks include: (1) Video-MME (Fu et al., 2025), with 2.7K questions over 900 videos ranging from 11 seconds to one hour across diverse domains; (2) TempCompass (Liu et al., 2024), with 7,540 instructions over 410 videos assessing action, direction, speed, event order, and attribute change; and (3) VideoEval-Pro (Ma et al., 2025), with 1,289 open-ended questions over 465 videos longer than ten minutes for realistic long-video understanding. The video reasoning benchmarks include: (4) MMVU (Zhao et al., 2025), with 3K expert-level questions across 27 disciplines requiring visual and domain-knowledge reasoning; (5) VSI-Bench (Yang et al., 2025), with over 5K questions from 288 egocentric indoor videos for spatial and spatiotemporal reasoning; (6) NExT-GQA (Xiao et al., 2024), with 8,911 causal and temporal questions paired with grounded temporal evidence; and (7) Video-MMMU (Hu et al., 2026), with 900 questions over 300 college-level educational videos assessing perception, comprehension, and knowledge transfer across six disciplines.

4.4 Baselines

To provide a broad comparison, we report results from proprietary models, including GPT-4o (Hurst et al., 2024) and Gemini 1.5 Pro (Team et al., 2024), as well as general-purpose open-source video MLLMs, including LLaMA-VID (Li et al., 2024b), LLaVA-OneVision-7B (Li et al., 2024a), Kangaroo-8B (Liu et al., 2026a), and VILA-1.5-8B (Lin et al., 2024). We further compare with recent reinforcement-learning-based video reasoning models, including Video-R1-7B (Feng et al., 2026), VideoRFT (Wang et al., 2026a), VideoChat-R1 (Wang et al., 2026b), and Time-R1-7B (Wang et al., 2026c). For controlled comparisons on both Qwen2.5-VL-7B and Qwen2-VL-7B backbones, we include the original instruction model, supervised fine-tuning on the same training set, and standard GRPO under the same experimental configuration, allowing us to isolate the contribution of the proposed frame-aware exploration and credit assignment.

4.5 Performance Analysis

Table 1 compares FrameGRPO with the baseline methods across general video understanding and video reasoning benchmarks. On the general benchmarks, FrameGRPO consistently improves upon standard GRPO for both backbones, with gains spanning comprehensive video question answering, fine-grained temporal understanding, and open-ended long-video evaluation. This shows that varying the frame sampling ratio during rollout collection improves temporal exploration without sacrificing the broad video-understanding ability of the underlying model.

Table 2: **Ablation Study of FrameGRPO Modules.** MFRoll and TempCred denote Multi-Frame Rollouts and Temporal Credit, respectively. All results are reported as accuracy (%).

Modules		Video General Benchmark			Video Reasoning Benchmark			
MFRoll	TempCred	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
✗	✗	60.0	74.4	41.0	65.6	34.1	79.5	52.0
✓	✗	60.7	74.4	41.8	66.3	34.7	**79.8**	52.2
✓	✓	**61.3**	**74.9**	**42.4**	**66.7**	**37.9**	79.6	**54.2**

Table 3: **Ablation Study of Policy Optimization Input.** We compare conditioning policy optimization only on the partial-ratio input with conditioning only on the full-ratio input. All results are reported as accuracy (%).

Optimization Input	Video General Benchmark			Video Reasoning Benchmark			
	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
Partial-Ratio Input	60.2	74.0	41.0	**66.7**	32.9	79.4	50.1
Full-Ratio Input (Ours)	**61.3**	**74.9**	**42.4**	**66.7**	**37.9**	**79.6**	**54.2**

FrameGRPO also exhibits strong performance on the reasoning benchmarks, which cover expert-level multidisciplinary reasoning, spatial and spatiotemporal understanding, temporally grounded question answering, and knowledge acquisition from educational videos. It surpasses GRPO on every benchmark with Qwen2.5-VL-7B and in nearly all settings with Qwen2-VL-7B, while achieving leading performance among similar-scale open-source methods on most benchmarks. The consistent trend across two model families suggests that the improvements are not specific to a single backbone. Moreover, the broad reasoning gains support our motivation that complementary temporal views can expose useful trajectories that are difficult to discover when all rollouts share a fixed visual context, while gated credit assignment effectively incorporates these discoveries into policy learning.

We further observe that supervised fine-tuning on the same training set does not provide consistent gains and often degrades the base models. In contrast, standard GRPO yields substantial improvements, which FrameGRPO further strengthens across both general understanding and reasoning tasks. This comparison indicates that the performance gains arise primarily from reinforcement learning and the proposed frame-aware exploration and credit assignment, rather than from exposure to the training data alone.

4.6 Ablation Study

Method Design. To examine the contribution of each component in FrameGRPO, we compare standard GRPO, GRPO augmented only with multi-frame rollouts, and the complete method combining multi-frame rollouts with Gated Temporal Efficiency Credit. As shown in Table 2, introducing multi-frame rollouts consistently improves performance across the evaluation suite, confirming that varying the visual context within a rollout group broadens temporal exploration and provides more informative group-relative supervision. Adding Gated Temporal Efficiency Credit yields further improvements on most tasks and produces the strongest overall performance. The consistent gains on most benchmarks demonstrate that the two components are complementary: multi-frame rollouts expose useful reasoning trajectories from alternative temporal views, while the credit mechanism selectively preserves and reinforces these discoveries when all rollouts in the fixed full-ratio reference subset fail.

Policy Optimization Input. FrameGRPO uses ratio-conditioned views for rollout exploration but conditions policy optimization only on the full-ratio input. To validate this design, we compare it with a variant that computes the policy and reference likelihoods using only the partial-ratio input associated with each rollout. As shown in Table 3, conditioning optimization on the full-ratio input matches or surpasses the partial-ratio alternative across all benchmarks, with particularly clear advantages on spatial reasoning and knowledge-acquisition tasks. These results indicate that partial-ratio views are valuable for discovering complementary trajectories, whereas reattaching those trajectories to the complete visual context during optimization more effectively transfers the discovered reasoning behavior into the policy.

Table 4: **Sensitivity Analysis of Hyperparameter α.** Results under varying α settings are reported on seven video benchmarks. The best performance for each benchmark is highlighted in bold. The configuration highlighted in blue corresponds to the setting used in the main results.

Setting	Video General Benchmark			Video Reasoning Benchmark			
	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
Vanilla GRPO	60.0	74.4	41.0	65.6	34.1	79.5	52.0
FrameGRPO with Different α							
$\alpha = 0.1$	**61.3**	**75.3**	**43.1**	**67.7**	35.8	**79.6**	52.6
$\alpha = 0.2$	**61.3**	74.9	42.4	66.7	**37.9**	**79.6**	**54.2**
$\alpha = 0.3$	61.0	75.0	41.5	65.4	34.3	**79.6**	52.0

Hyperparameter Sensitivity. We investigate the sensitivity of FrameGRPO to α, which controls the strength of Gated Temporal Efficiency Credit. As shown in Table 4, FrameGRPO retains improvements over standard GRPO on most benchmarks across the tested range, demonstrating that its effectiveness is robust to the choice of α. A smaller credit weight performs favorably on several general and expert-level reasoning tasks, whereas the default moderate setting provides the strongest overall balance, with particularly clear benefits on challenging spatial and knowledge-acquisition tasks. Increasing the weight further weakens performance on some tasks, as an excessive credit may overemphasize lower-ratio successes relative to the standard group-relative learning signal. Overall, FrameGRPO does not require precise tuning of α, while a moderate value offers a reliable trade-off across diverse video understanding and reasoning settings.

5 Limitation

FrameGRPO derives its additional training signal from differences among temporally aligned views, and its benefit therefore depends on how strongly a question responds to changes in frame density. For appearance-centric questions that can be answered from a few isolated frames, varying the sampling ratio may leave the rollout distribution nearly unchanged and provide little extra reward diversity. At the other extreme, aggressive subsampling can omit brief but decisive events when precise temporal localization is required. A promising direction is to estimate the informativeness and consistency of each temporal view and adapt both view construction and cross-view credit to individual video–question pairs.

6 Conclusion

In this work, we developed **FrameGRPO** to improve reinforcement learning for video reasoning by extending policy exploration from linguistic trajectories to temporal evidence. Its Multi-Frame Rollouts reuse a maximal-frame decode to construct temporally aligned views at different sampling ratios, enabling more informative cross-view comparisons within each GRPO group. To exploit useful discoveries from sparse views without systematically favoring fewer frames, Gated Temporal Efficiency Credit selectively strengthens a correct lower-ratio trajectory only when all rollouts in the fixed full-ratio reference subset are unsuccessful. Full-ratio re-conditioning further transfers these discoveries to a unified policy while requiring no additional mechanism at inference time. Extensive evaluations across diverse model backbones and video understanding and reasoning benchmarks demonstrate that FrameGRPO delivers robust overall improvements over standard GRPO. Together with the greater within-group reward variation observed during training, these results highlight temporal-view diversification as an effective approach to exploration and credit assignment for video MLLMs.

AI use statement

In this work, we used generative AI tools only to assist with drafting and polishing the manuscript for clarity, grammar, and readability. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design the research methodology or experiments, implement methods, translate content, clean or reformat datasets, conduct qualitative or thematic data analysis,

or interpret results. All AI-assisted text was manually reviewed, revised, and verified by the authors for accuracy, originality, and consistency with the research. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. *Advances in Neural Information Processing Systems*, 38:143018–143046, 2026.

Emad Bahrami, Olga Zatsarynna, Parth Pathak, Sunando Sengupta, Juergen Gall, and Mohsen Fayyaz. Strive: Structured spatiotemporal exploration for reinforcement learning in video question answering. *arXiv preprint arXiv:2604.01824*, 2026.

Zongsheng Cao, Anran Liu, Jun Xie, Feng Chen, Lang Chen, Jing Li, Zhepeng Wang, and Zigan Wang. Videotrace-r1: Long video-based retrieval-augmented generation via reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism. *arXiv preprint arXiv:2603.29252*, 2026a.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870, 2026b.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensesearch: Empowering vision-language models with high-resolution agentic search-reasoning via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26508–26517, 2026.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. Appo: Attention-guided perception policy optimization for video reasoning. *arXiv preprint arXiv:2602.23823*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Lin Fu, Zheyuan Yang, Yang Wang, Tingyu Song, Arman Cohan, and Yilun Zhao. Videokr: Towards knowledge-and reasoning-intensive video understanding. *arXiv preprint arXiv:2606.05259*, 2026.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage. *arXiv preprint arXiv:2603.01106*, 2026.

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

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *arXiv preprint arXiv:2605.21125*, 2026a.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, volume 2026, pp. 74904–74933, 2026b.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jitesh Jain, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, Humphrey Shi, et al. Sage: Training smart any-horizon agents for long video reasoning with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41478–41488, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026a.

Xuchen Li, Xuzhao Li, Shiyu Hu, and Kaiqi Huang. Select less, reason more: Prioritizing evidence purity for video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25621–25632, 2026b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European conference on computer vision*, pp. 323–340. Springer, 2024b.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. *Advances in Neural Information Processing Systems*, 38:40692–40716, 2026c.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26679–26689. IEEE, 2024.

Xiongkun Linghu, Jiangyong Huang, Baoxiong Jia, and Siyuan Huang. 3d-rft: Reinforcement fine-tuning for video-based 3d scene understanding. *arXiv preprint arXiv:2603.04976*, 2026.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input: J. liu et al. *International Journal of Computer Vision*, 134(3):114, 2026a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 94069–94086, 2026b.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8376–8405, 2026.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *Advances in Neural Information Processing Systems*, 38:138605–138632, 2026.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16944–16953, 2026.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *Advances in Neural Information Processing Systems*, 38:150977–151013, 2026.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026a.

Xiaodong Wang, Zhirong Wu, Langling Huang, Yuxi Zheng, and Peixi Peng. Incentivizing versatile video reasoning in mllms via data-efficient reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5444–5454, 2026b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026c.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Jiang Xudong. Video-ktr: Reinforcing video reasoning via key token attribution. In *International Conference on Learning Representations*, volume 2026, pp. 71553–71570, 2026d.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. In *International Conference on Learning Representations*, volume 2026, pp. 78298–78319, 2026a.

Peiran Wu, Zhuorui Yu, Yunze Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. Marc: Memory-augmented rl token compression for efficient video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 13120–13135, 2026b.

Tao Wu, Li Yang, Gen Zhan, Yabin Zhang, Yiting Liao, Junlin Li, Deliang Fu, Li Zhang, and Limin Wang. Tempr1: Improving temporal understanding of mllms via temporal-aware multi-task reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2756–2767, 2026c.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

Wenyi Xiao and Leilei Gan. Fast-slow thinking grpo for large vision-language model reasoning. *Advances in Neural Information Processing Systems*, 38:171601–171631, 2026.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. In *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capabilities of multimodal large language models via share-grpo. *Advances in Neural Information Processing Systems*, 38: 80424–80451, 2026.

Hee Suk Yoon, Eunseop Yoon, Ji Woo Hong, SooHwan Eom, Gwanhyeong Koo, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Pdcr: Perception-decomposed confidence reward for vision-language reasoning. *arXiv preprint arXiv:2605.13467*, 2026.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:94827–94853, 2026a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026b.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026a.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026b.

Xichen Zhang, Sitong Wu, Yinghao Zhu, Haoru Tan, Shaozuo Yu, Ziyi He, and Jiaya Jia. Scaf-grpo: Scaffolded group relative policy optimization for enhancing llm reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 131946–131974, 2026c.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489. IEEE, 2025.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, et al. Active-o3: Empowering mllms with active perception via pure reinforcement learning. *arXiv preprint arXiv:2505.21457*, 2025.

Zirui Zhu, Hailun Xu, Yang Luo, LIU YONG, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 98463–98486, 2026.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. Videobrain: Learning adaptive frame sampling for long video understanding. *arXiv preprint arXiv:2602.04094*, 2026.

A Additional Ablation Studies And Results

Robustness Across Random Seeds. To evaluate robustness to training stochasticity, we repeat GRPO and FrameGRPO with three random seeds on Qwen2.5-VL-7B while keeping all other settings fixed. As shown in Table 5, FrameGRPO consistently outperforms the corresponding GRPO baseline across nearly the entire evaluation suite for every seed. This consistent trend indicates that the benefits of multi-frame exploration and gated temporal credit are robust to initialization and data ordering rather than arising from a favorable individual run.

Table 5: **Robustness Across Random Seeds.** Results for GRPO and FrameGRPO with Qwen2.5-VL-7B under three random seeds. All results are reported as accuracy (%). The best result within each seed is highlighted in bold.

Configuration		Video General Benchmark			Video Reasoning Benchmark			
Method	Seed	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
GRPO	42	60.0	74.4	41.0	65.6	34.1	79.5	52.0
FrameGRPO		**61.3**	**74.9**	**42.4**	**66.7**	**37.9**	**79.6**	**54.2**
GRPO	43	59.8	73.8	39.4	66.4	34.0	**80.0**	50.2
FrameGRPO		**61.3**	**75.0**	**42.1**	**68.0**	**35.1**	79.8	**51.4**
GRPO	44	60.2	73.8	41.8	65.6	31.7	79.4	52.4
FrameGRPO		**61.5**	**74.9**	**42.2**	**67.7**	**34.3**	**79.6**	**52.7**

Scalability Analysis. We further compare GRPO and FrameGRPO using training subsets of different sizes while keeping the remaining settings fixed. As shown in Table 6, FrameGRPO consistently improves the corresponding GRPO baseline across nearly all evaluation settings, from the smallest subset to the full training set. The gains remain evident in the low-data regime and become more broadly consistent as additional training data are introduced, indicating that FrameGRPO is effective across different data budgets and can make productive use of increased supervision.

Table 6: **Robustness Across Training-Set Sizes.** Results for GRPO and FrameGRPO with Qwen2.5-VL-7B using different numbers of training examples. All results are reported as accuracy (%). The best result within each data size is highlighted in bold.

Configuration		Video General Benchmark			Video Reasoning Benchmark			
Method	Data Size	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
GRPO	2K	59.6	73.7	41.9	66.2	31.0	**79.7**	51.6
FrameGRPO		**60.6**	**74.0**	**42.6**	**66.9**	**33.1**	78.9	**51.7**
GRPO	4K	60.6	74.0	42.5	65.3	33.4	**79.6**	51.2
FrameGRPO		**61.4**	**75.2**	**42.6**	**66.7**	**37.6**	79.5	**52.4**
GRPO	8K	60.0	74.4	41.0	65.6	34.1	79.5	52.0
FrameGRPO		**61.3**	**74.9**	**42.4**	**66.7**	**37.9**	**79.6**	**54.2**

Temporal Efficiency Credit Design. To isolate the importance of ratio-aware weighting and selective gating in Temporal Efficiency Credit (TEC), we compare a uniform credit with a fixed weight, ratio-aware TEC applied without the full-ratio gate, and our gated TEC under the same multi-frame rollout setting. As shown in Table 7, assigning a uniform bonus or applying ratio-aware credit without gating does not yield consistent improvements across the evaluation suite. A uniform bonus ignores both the amount of temporal evidence used by a trajectory and whether its success provides

information beyond the full-ratio reference, whereas ungated TEC introduces an unconditional preference for successful lower-ratio trajectories even when the full-ratio rollouts already provide valid learning signals. Our gated design achieves the strongest and most consistent overall performance by preserving the standard group-relative objective whenever the full-ratio reference succeeds and activating additional credit only when a lower-ratio trajectory reveals a complementary success. This comparison confirms that the benefit of TEC arises not merely from favoring temporally efficient trajectories, but from selectively crediting them when they contribute otherwise undiscovered reasoning evidence.

Table 7: **Ablation Study of Temporal Efficiency Credit.** We compare a uniform credit with a fixed weight of 0.2, TEC without the full-ratio gate, and the proposed gated TEC. All results are reported as accuracy (%). The best result for each benchmark is highlighted in bold.

Credit Design	Video General Benchmark			Video Reasoning Benchmark			
	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
Uniform Credit (0.2)	60.7	**74.9**	42.0	65.1	34.7	79.3	51.3
TEC w/o Gate	60.1	74.6	41.1	65.9	34.9	79.5	51.9
Gated TEC (Ours)	**61.3**	**74.9**	**42.4**	**66.7**	**37.9**	**79.6**	**54.2**

Full-Ratio Reference Subset Size. We study the effect of the number of full-ratio reference rollouts $K_\star$ while keeping the group size fixed. As shown in Table 8, using a single reference rollout makes the gate overly sensitive to the outcome of one stochastic generation, leading to a less reliable estimate of whether the full-ratio view has failed. Conversely, a large reference subset reserves too many rollout slots for the same full-ratio view and restricts exploration across alternative temporal contexts. We therefore adopt $K_\star = 2$, which provides a more reliable gating signal while retaining sufficient rollout capacity for temporal exploration.

Table 8: **Ablation Study of Full-Ratio Reference Subset Size.** We vary $K_\star$ while keeping the group size fixed. All results are reported as accuracy (%). The best result for each benchmark is highlighted in bold, and the configuration used in our main experiments is highlighted in blue.

Subset Size	Video General Benchmark			Video Reasoning Benchmark			
	Video-MME	TempCompass	VideoEval-Pro	MMVU	VSI-Bench	NExT-GQA	Video-MMMU
$K_\star = 1$	60.7	74.6	41.1	65.1	35.3	78.0	49.3
$K_\star = 2$	**61.3**	74.9	**42.4**	**66.7**	**37.9**	**79.6**	**54.2**
$K_\star = 4$	60.6	**75.0**	41.4	65.0	36.6	78.9	53.4

Training Efficiency. We compare the end-to-end training cost of GRPO and FrameGRPO in Table 9. FrameGRPO maintains comparable training cost on Qwen2.5-VL-7B and incurs only moderate runtime overhead on Qwen2-VL-7B, while substantially reducing peak GPU memory for both backbones. These results show that multi-frame exploration and gated credit assignment can be integrated without increasing memory demand and with limited additional computation.

Table 9: **Training Efficiency Comparison.** All runs use eight NVIDIA H20 GPUs. GPU-hours are aggregated across all GPUs.

Method	Wall-Clock Time	GPU-Hours	Peak GPU Memory
Qwen2.5-VL-7B			
GRPO	21h 24min	171.2	94.95 GiB
FrameGRPO	21h 19min	170.5	57.46 GiB
Qwen2-VL-7B			
GRPO	17h 01min	136.1	94.63 GiB
FrameGRPO	19h 51min	158.8	63.80 GiB

B More Implementation Details

B.1 Training Details

Table 10 summarizes the training configuration used in our FrameGRPO experiments.

Table 10: **Training Hyperparameters for FrameGRPO.** The table reports the default configuration used in our experiments.

Hyperparameter	Value
Sampling	
Group size G	8 rollouts per prompt
Training sampling temperature	1.0
Maximum prompt / completion length	8,192 / 768 tokens
Optimization	
Optimizer	AdamW
Learning rate	1×10^{-6}
Gradient clipping norm	1.0
Training epochs / optimizer steps	1 / 1,000
Prompt batch size per GPU	1
GRPO / FrameGRPO	
KL coefficient β	0.04
Full-ratio frame count $f_\star$	32
Full-ratio reference subset size $K_\star$	2
TEC coefficient α	0.2

B.2 Prompt Template

We use the same prompt template for both training and inference, as shown in Figure 4. The question and task-specific output format are inserted into the corresponding placeholders, ensuring that the reasoning instructions and answer format remain consistent across optimization and evaluation.

```
{Question}
Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags.
{OutputTemplate}

OutputTemplate:
"multiple choice": Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.
"regression": " Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.
```

Figure 4: **Prompt template used for both training and inference.**

C Reasoning Example

We provide two representative case studies comparing FrameGRPO with vanilla GRPO to highlight the improvements in video perception and reasoning, as shown in Figures 5 and 6, respectively.

Question: Why did the boy stretch out his hand halfway ?
A. touch woman hand
B. to get the toy
C. for hand gesture with someone
D. take the camera
E. replying to someone



FrameGRPO:

<think>
The boy's action of stretching out his hand halfway suggests a desire to interact with something or someone. Given the context, it's logical to consider the most likely reason. The boy is walking on a sidewalk and approaches a woman who is holding a toy. The boy's gesture could be interpreted as an attempt to reach for the toy. This is a common behavior for children when they see something they want to play with. Let me think through the options:
- A. touch woman hand: This seems less likely because the boy is not reaching for the woman's hand directly.
- B. to get the toy: This fits the context and the boy's typical behavior.
- C. for hand gesture with someone: While possible, it's less likely given the context.
- D. take the camera: The boy is not reaching for the camera.
- E. replying to someone: The boy is not engaging in a conversation.
</think>
<answer>B</answer>

GRPO:

<think>
The boy's action of stretching out his hand halfway could be interpreted in several ways, but given the context of the video, it seems most likely that he is reaching out to touch the woman's hand. This is a common gesture for a child to show affection or to seek guidance. The woman is walking alongside him, and it's natural for a child to want to hold hands with an adult for safety and comfort. The other options, such as getting a toy or replying to someone, are less likely in this scenario as there is no indication of a toy being present or someone else being involved.
</think>
<answer>A</answer>

Figure 5: **Reasoning example on NExT-GQA.**

Question: Which food item is cut in half with a knife by the person in the video?
A. Cucumber. B. Tomato.
C. Tofu. D. Almond.



FrameGRPO:

<think>
Let me think about this. The video shows a person cutting various items, but the question specifically asks about an item cut in half with a knife. I see the person cutting a cucumber in half, a tomato in half, and a block of tofu. However, the question is about an item cut in half with a knife. The cucumber is cut in half, but it's not specified as being cut with a knife. The tomato is cut in half with a knife, but the question asks for an item cut in half with a knife. The tofu is cut into smaller pieces, not in half. So, the correct answer should be the item that is cut in half with a knife.
</think>
<answer>B</answer>

GRPO:

<think>
Let me think about this. The video shows a person cutting a cucumber in half with a knife. The cucumber is clearly visible and is being cut in half, which matches the description in the question. The other options like tomato, tofu, and almond are not being cut in half in the video. So, the correct answer should be related to the cucumber.
</think>
<answer>A</answer>

Figure 6: **Reasoning example on Video-MME.**