Dual-Group Exploration: Reinforcement Learning for Video Understanding in Multimodal Large Language Models

Qucheng Peng*, **Rui Hou**[1], **Jingyi Zhang**[1], **Shengjie Zhu**[1], **Yingfu Xu**[1], **Haotian Xu**[1], **Noranart Vesdapunt**[1], **Chen Chen**[2], **Ning Zhou**[1]
[1]Amazon
[2]Center for Research in Computer Vision, University of Central Florida
pengdavid2025@gmail.com, {shngjz, atxu, solves}@amazon.com
yingfuxu@amazon.nl, zhjingyi@lab126.com, chen.chen@crcv.ucf.edu

Abstract

Reinforcement Learning (RL) has significantly enhanced reasoning in Large Language Models. However, applying state-of-the-art RL methods like Group Relative Policy Optimization (GRPO) to video-based Multimodal LLMs (MLLMs) remains sub-optimal. We identify two key exploration bottlenecks: (1) **Advantage Vanishing**: generating multiple responses from invariant video contexts leads to near-uniform rewards, neutralizing learning signals after in-group standardization; (2) **Visual Overfitting**: standard RL lacks mechanisms to regularize sensitivity to visual perturbations under single-epoch training, causing policies to overfit to sparse temporal cues. To address these, we propose Dual-Group Exploration, a framework revitalizing learning signals at both input and output levels. At the *input level*, a Dual-Group Sampling strategy breaks visual determinism via complementary temporal sequences. At the *output level*, an Entropy-rated ELO Advantage captures inter-group quality variances. Concurrently, a Cross-Visual Counterfactual Weight establishes a Visual Consistency Region, safely driving exploration by enforcing reasoning consistency across varied views. Achieving state-of-the-art performance across six video reasoning benchmarks, our method demonstrates that consistency-guided exploration is beneficial for video understanding.

1 Introduction

Recent breakthroughs in Reinforcement Learning (RL) have catalyzed significant advancements in the reasoning capabilities of Large Language Models (LLMs) (Ghasemi and Ebrahimi, 2024; Ouyang et al., 2022; Shao et al., 2024). A cornerstone of this progress is Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which, as exemplified by DeepSeek-R1 (Guo et al., 2025), demonstrates that rule-based RL can elicit emergent and robust reasoning through efficient in-group advantage estimation. Building on this success, recent efforts have extended GRPO to Multimodal LLMs (MLLMs) (Wang et al., 2025; Bai et al., 2025b; Xiaomi, 2025) and have achieved substantial gains in image-text reasoning. However, when applied to *video understanding*, the performance gains of GRPO are notably diminished compared to text-based or image-centric tasks (Feng et al., 2025; Chen et al., 2025a). We identify that this limitation is not merely a data issue but stems from two key bottlenecks in existing RL paradigms: the invariant visual contexts causing the RL exploration signal to collapse, and the lack of visual regularization under single-epoch training constraints.

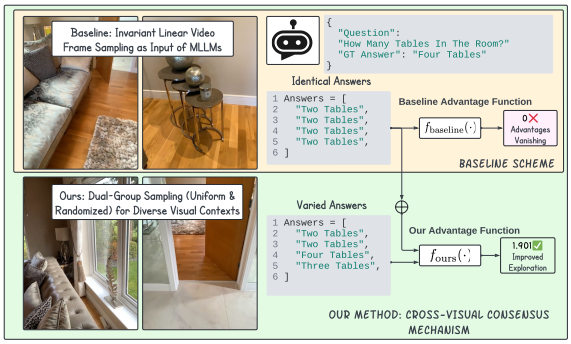


Figure 1: Overcoming Distributional Stagnation via Dual-Group Exploration. **Upper:** Existing RL methods rely on identical linear video sampling, leading to collapsed response diversity. This causes the *advantage vanishing* problem, where uniform rewards yield zero gradients after group standardization, effectively stalling the policy update. **Below:** Our method introduces Dual-Group Sampling (Uniform vs. Randomized) to elicit varied reasoning trajectories from the same video. By leveraging an ***Entropy-rated ELO*** reward function, we identify cross-visual consensus to generate meaningful advantage signals (1.901 vs. 0).

Through careful analysis, we identify that the first exploration bottleneck manifests as a **vanishing advantage problem** rooted in *invariant video*

*This work was completed during Qucheng Peng's internship at Amazon.

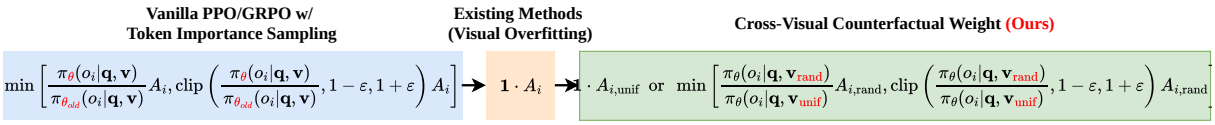


Figure 2: Evolution of Consistency Region Mechanisms for Video RL. **Vanilla PPO/GRPO** (Left) utilizes an importance ratio ($\pi_\theta/\pi_{\theta_{\text{old}}}$) to constrain policy drift, which is computationally prohibitive for single-epoch video MLLMs. **Existing Methods** (Middle) force this ratio to a constant **1**, stripping the policy of its adaptive protections and leading to *visual overfitting*. **Our Method** (Right) introduces a **Cross-Visual Counterfactual Weight** ($\pi_\theta(\cdot|\mathbf{v}_{\text{rand}})/\pi_\theta(\cdot|\mathbf{v}_{\text{unif}})$) to establish a *Visual Consistency Region (VCR)*. This safely empowers bold exploration by ensuring cross-visual consistency, requiring no additional multi-epoch overhead.

contexts. As illustrated in Fig. 1 (top), current RL paradigms for video typically employ fixed sampling strategies, such as uniform temporal sampling, and repeat the identical frame set as input for all responses within a GRPO group. This deterministic input constraint leads to a **numerical collapse** of learning signals. Specifically, when an MLLM receives identical video frames, it tends to generate highly homogeneous outputs (*e.g.*, Answer: 2, 2, 2, 2") with near-uniform rewards (*e.g.*, Reward: 0.5, 0.5, 0.5, 0.5"). After in-group standardization, the resulting advantage vector $\mathbf{A}$ collapses toward zero ($\mathbf{A} \approx \mathbf{0}$), effectively neutralizing the gradient and depriving the model of the discriminative signal needed to distinguish superior trajectories. While some recent works (Feng et al., 2025) attempt to introduce temporal reward thresholds via shuffled frames, they still compute advantages based on static inputs, failing to address the fundamental neutralization of the exploration signal. In contrast, our approach diversifies the input space as shown in Fig. 1 (bottom), providing varied visual perspectives through extra sampling encourages the model to explore the reasoning manifold, leading to differentiated rewards (*e.g.*, "Reward: 0.5, 0.5, 1, 0.75") and meaningful, non-zero advantages that provide a clear directional signal for policy improvement.

Beyond the structural limitations of advantage estimation, the learning process is further hampered by **visual overfitting** under single-epoch constraints. In standard PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024), the Importance Sampling (IS) ratio ($\frac{\pi_\theta}{\pi_{\theta_{old}}}$) serves as a critical adaptive mechanism to constrain temporal policy drift, providing a safe boundary for exploration across multiple optimization epochs. However, due to the prohibitive memory footprint of processing high-resolution video frames, video MLLMs (Feng et al., 2025; Chen et al., 2025a,b) are generally restricted to single-epoch optimization, where the traditional IS ratio inherently degenerates to **1**. Without this trust-consistency protection, the model becomes hypersensitive to specific visual contexts and overfits to them. Rather than attempting to resurrect the defunct IS ratio, we identify the need for a spatial-temporal alternative to safely anchor the policy, thereby empowering the model to explore diverse reasoning trajectories without losing consistency.

To address these bottlenecks, we propose ***Dual-Group Exploration***, a unified RL framework revitalizing learning signals at both input and output levels. At the *input* level, a **Dual-Group Sampling** strategy breaks visual determinism by generating complementary (uniform and randomized) video perspectives. At the *output* level, we restore discriminative learning via two mechanisms. First, an **Entropy-Rated ELO** (Christiano et al., 2017) advantage treats cross-context responses as *competing agents*, assigning high rewards strictly to robust, high-confidence reasoning. Second, a **Cross-Visual Counterfactual Weight** serves as a cross-visual analogue of adaptive importance weighting. By measuring distributional divergence across visual contexts, it establishes a *Visual Consistency Region (VCR)*. Ultimately, input-level perturbations drive *exploration*, while the VCR acts as a steering wheel, encouraging the policy to boldly explore reasoning pathways without losing logical consistency. Our main contributions are summarized as follows:

- We propose a Dual-Group Sampling strategy to drive input-level exploration. By applying complementary video perspectives, it breaks visual determinism and directly mitigates the *advantage vanishing* bottleneck.
- We restore output-level discriminative learning via two mechanisms: an Entropy-rated ELO Advantage that treats responses as *competing agents*

to reward robust reasoning, and a Cross-Visual Counterfactual Weight that establishes a visual consistency region for safe exploration.
- We achieve state-of-the-art performance across six video benchmarks while reducing the RL training time by 33%, proving the effectiveness of exploration.

2 Related Works

LLM Reasoning via RL. Transitioning from dense supervision (Gao et al., 2024; Li and Li, 2024), recent advancements leverage rule-based reinforcement learning (RL) to unlock emergent reasoning. Notably, DeepSeek-R1 (Guo et al., 2025) demonstrates that outcome-only rewards can elicit robust, long-form chain-of-thought (CoT) trajectories. This paradigm has been rapidly adapted across textual and visual domains (Hu et al., 2025a; Team et al., 2025; Huang et al., 2025; Xie et al., 2025). However, directly lifting text-centric RL pipelines to video reasoning introduces distinct bottlenecks, primarily due to catastrophic training overhead and sparse verification signals.

Video MLLMs and Reinforcement Learning. Traditional video-based Multimodal LLMs (MLLMs) focus heavily on architectural scaling or token efficiency to capture spatial-temporal dynamics (Cheng et al., 2024; Zhang et al., 2024b; Chen et al., 2024; Li et al., 2024c; Zhang et al., 2024a; Zou et al., 2026). More recently, pioneering works have begun integrating RL into video understanding. For instance, Video-R1 (Feng et al., 2025) introduces T-GRPO to explicitly promote temporal reasoning steps, while GRPO-CARE (Chen et al., 2025a) optimizes the intra-group alignment via adaptive bonuses. Distinct from these methods that optimize policies over static visual inputs under fixed data trajectories, our framework targets the fundamental gradient collapse by reshaping the exploration manifold at both the input context and policy optimization levels.

3 Methodology

3.1 Preliminaries and Motivations

GRPO Baseline. Given a video understanding dataset $\mathcal{D} = \{(v^i, q^i, a^i)\}_{i=1}^N$, a multi-modal policy π_θ generates a group of G outputs $\mathcal{P}^i = \{p_g^i\}_{g=1}^G$ for each sample. Following GRPO (Shao et al., 2024; Guo et al., 2025), the standardized advantage A_g^i for each response is calculated as:

$$A_g^i = \frac{r_g^i - \text{mean}(\mathbf{r}^i)}{\text{std}(\mathbf{r}^i) + \delta}, \tag{1}$$

where $\mathbf{r}^i = \{r_1^i, \ldots, r_G^i\}$ represents the reward vector, and $\delta = 10^{-4}$ handles numerical stability.

Due to the prohibitive computational cost of temporal modeling, existing video RL frameworks like Video-R1 (Feng et al., 2025) restrict policy updates to a single epoch, effectively fixing the token-level importance sampling ratio to **1**. The simplified optimization objective is formalized as:

$$J_{\text{GRPO-VU}}(\theta) = \mathbb{E}_{\mathcal{D},\mathcal{P}^i\sim\pi_\theta}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|p_g^i|}\sum_{t=1}^{|p_g^i|}A_{g,t}^i\right] - \alpha\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\theta_{\text{ref}}}), \tag{2}$$

where $A_{g,t}^i$ denotes the token-level advantage mapped from A_g^i, and α regularizes deviations from the reference policy $\pi_{\theta_{\text{ref}}}$.

Exploration Bottlenecks. Despite its efficiency, this paradigm suffers from a severe dual exploration crisis in video tasks: (1) *Advantage Vanishing:* Relying on static, invariant video frame inputs across the group inherently collapses output variance. Highly homogeneous responses yield near-uniform rewards, forcefully driving $A_g^i \to 0$ after intra-group standardization and neutralizing the critical gradient signals needed for policy improvement. (2) *Visual Overfitting:* The computational constraint of single-epoch training forces the importance sampling ratio to 1. This strips the policy of its trust-consistency protections, causing the model to become hypersensitive to specific temporal cues and overfit to the fixed visual context.

To mitigate these visual and policy-level bottlenecks, we introduce ***Dual-Group Exploration*** (Fig. 3), a unified framework that revitalizes learning signals via input-level dual-group sampling, confidence-aware entropy-rated ELO advantages, and cross-visual counterfactual weighting.

3.2 Framework

Dual-group Video Frame Sampling. To mitigate the *exploration bottleneck* where repeated visual contexts lead to collapsed policy gradients, we propose a **dual-group sampling** strategy. Traditional methods employ a fixed uniform sampling interval (*e.g.*, selecting frames at positions $\{1, 50, 100, 150\}$ from a 200-frame video with

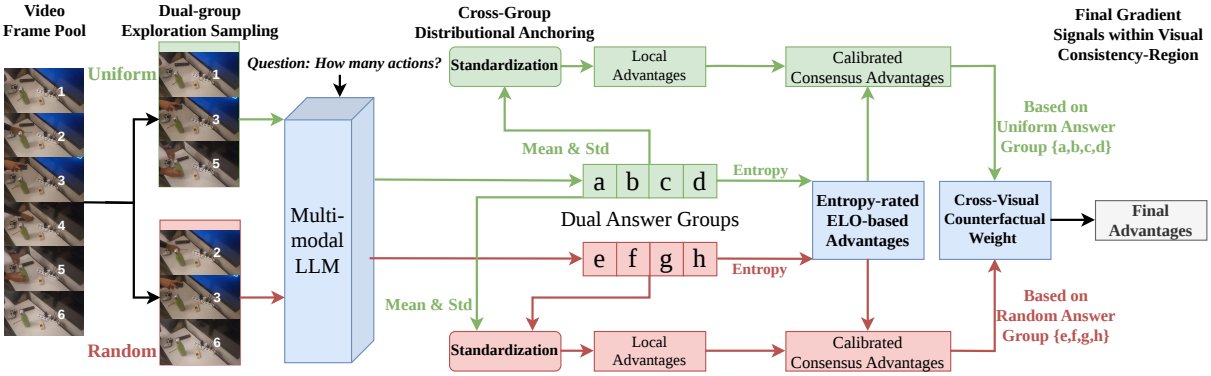


Figure 3: Overview of the Dual-Group Exploration Framework for Video Reasoning. Our framework utilizes a dual-path architecture to overcome *advantage vanishing* and *visual overfitting*. Videos are processed through Dual-Group Sampling: a uniform sampling path (green) providing stable temporal anchors, and a random sampling path (chestnut) for exploratory diversity. Both groups generate reasoning traces via a Multimodal LLM, which are processed through Cross-Group Distributional Anchoring to yield local advantages. To mitigate "confident-but-incorrect" hallucinations, we introduce Entropy-rated ELO-based Advantages to identify cross-visual consensus. Finally, a **Cross-Visual Counterfactual Weight** establishes a *Visual Consistency Region (VCR)*, merging both paths to produce robust gradient signals that empower safe, consistency-guided policy exploration.

$F = 4$). While it ensures temporal coverage, it also results in *near-uniform rewards* across a response group, effectively neutralizing the advantage signal after the standardization operation.

Our approach treats different frame subsets as *complementary visual perspectives* of the same semantic content. We first construct an expanded candidate pool by densely sampling $k \times F$ frames ($k \in \{2, 3, 4\}$). From this pool, we derive two distinct input configurations:

- **Uniform group** ($v^i_{\textbf{unif}}$): Maintains the standard temporal structure to provide a stable baseline for policy updates.
- **Random group** ($v^i_{\textbf{rand}}$): Introduces stochastic visual perturbations while preserving chronological order, thereby forcing MLLM to explore reasoning space under varied evidential contexts.

We set $G_{\text{unif}} = G_{\text{rand}} = G/2$ to ensure computational parity with existing GRPO implementations. This dual-group strategy enables the framework to capture sparse, non-periodic key events across extended temporal horizons by coupling deterministic coverage with randomized exploration, ensuring robustness in "needle-in-a-haystack" video scenarios.

Cross-Group Distributional Anchoring. Direct standardization within independent groups can lead to misaligned advantage scales and distributional shifts. To establish a common reference frame, we compute the advantage for the uniform group $A^i_{g,\text{unif}}$ using its in-group statistics:

$$A^i_{g,\text{unif}} = \frac{r^i_{g,\text{unif}} - \text{mean}(\mathbf{r}^i_{\text{unif}})}{\text{std}(\mathbf{r}^i_{\text{unif}}) + \delta}, \tag{3}$$

where $\mathbf{r}^i_{\text{unif}} = \{r^i_{1,\text{unif}}, r^i_{2,\text{unif}}, \ldots, r^i_{G_{\text{unif}},\text{unif}}\}$ denotes the reward vector for the uniformly sampled group. For the random group, we **anchor** its advantage estimation to the uniform group's reward distribution:

$$A^i_{g,\text{rand}} = \frac{r^i_{g,\text{rand}} - \text{mean}(\mathbf{r}^i_{\text{unif}})}{\text{std}(\mathbf{r}^i_{\text{unif}}) + \delta}. \tag{4}$$

We prioritize the uniform group as the **reference distribution** because its deterministic sampling offers the most consistent and stable coverage of the temporal domain. By utilizing its statistics as a *low-variance baseline*, we ensure that advantage estimates for exploratory trajectories are measured against a stable, representative anchor, thereby preventing erratic updates driven by the high stochasticity of randomized frame subsets.

Intrinsic Motivation via Entropy-rated ELO. To mitigate the *advantage vanishing*, we introduce an **intrinsic reward** based on the ELO rating system (Christiano et al., 2017). We utilize token-level entropy H^i_y as a proxy for *predictive confidence*:

$$H^i_y = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \exp(\ell^{(k)}_t) \cdot \ell^{(k)}_t, \tag{5}$$

where H^i_y reflects the model's *length-invariant reasoning certainty*. To mitigate the risk of rewarding "confident but incorrect" hallucinations, we

treat responses from different video frame groups as **competing agents**. Low entropy under a single view may reflect overconfident hallucination, whereas consistent low entropy across perturbed views better indicates grounded reasoning. This formulation functions as a cross-visual calibration check: by mapping entropy to a quality scale Q via standardization:

$$Q^i_{y,\text{unif}} = \mu + \sigma \cdot \frac{H^i_{y,\text{unif}} - \text{mean}(\mathbf{H}^i_{\text{unif}})}{\text{std}(\mathbf{H}^i_{\text{unif}})},$$
$$Q^i_{y,\text{rand}} = \mu + \sigma \cdot \frac{H^i_{y,\text{rand}} - \text{mean}(\mathbf{H}^i_{\text{rand}})}{\text{std}(\mathbf{H}^i_{\text{rand}})}, \quad (6)$$

where $\mathbf{H}^i_{\text{unif}} = \{H^i_{1,\text{unif}}, H^i_{2,\text{unif}}, \ldots, H^i_{G_{\text{unif}},\text{unif}}\}$ and $\mathbf{H}^i_{\text{rand}} = \{H^i_{1,\text{rand}}, H^i_{2,\text{rand}}, \ldots, H^i_{G_{\text{rand}},\text{rand}}\}$ denote the entropy vector for the uniformly and randomly sampled groups respectively. Besides, μ defines the center of the rating scale and σ controls the rating spread, ensuring numerical stability. Based on the quality scores, the expected scores are exhibited as:

$$E^i_{g,\text{unif}} = \frac{1}{T}\frac{1}{G_{\text{rand}}}\sum_{n=1}^{G_{\text{rand}}}\frac{1}{1+10^{(Q^i_{n,\text{rand}}-Q^i_{g,\text{unif}})/400}},$$
$$E^i_{g,\text{rand}} = \frac{1}{T}\frac{1}{G_{\text{unif}}}\sum_{m=1}^{G_{\text{unif}}}\frac{1}{1+10^{(Q^i_{m,\text{unif}}-Q^i_{g,\text{rand}})/400}}, \quad (7)$$

where the scaling factor of 400 follows the standard ELO-rating system (Christiano et al., 2017). By combining these with basic advantages, the enhanced advantages are represented as:

$$\hat{A}^i_{g,\text{unif}} = A^i_{g,\text{unif}} + \eta E^i_{g,\text{unif}}, \quad \hat{A}^i_{g,\text{rand}} = A^i_{g,\text{rand}} + \eta E^i_{g,\text{rand}}, \quad (8)$$

where η is a trade-off parameter between two types of advantages. This mechanism ensures that high rewards are only sustained by responses achieving cross-visual consensus for those that remain certain and consistent across varied visual samplings, thereby penalizing stochastic, over-confident hallucinations that fail under visual perturbations.

Cross-Visual Counterfactual Weighting. Standard GRPO uses the Importance Sampling (IS) ratio ($\pi_\theta/\pi_{\theta_{old}}$) to regulate temporal policy drift across multiple epochs (Shao et al., 2024). However, video RL is typically restricted to a single epoch due to prohibitive memory costs, forcing this temporal ratio to **1** and rendering the policy vulnerable to *visual overfitting*. To introduce adaptive regularization without multi-epoch overhead, we repurpose the IS mathematical structure into a **cross-visual counterfactual weight** $r^i_{g,t}$, anchoring the exploratory distribution against a stable counterfactual view. Formally, we treat the uniform group v^i_{unif} as the baseline anchor context and the randomized group v^i_{rand} as the exploratory context. The uniform baseline objective is:

$$J_{\text{unif}}(\theta) = \mathbb{E}_{\mathcal{D},\{p^i_g\}\sim\pi_\theta}\left[\frac{1}{G_{\text{unif}}}\sum_{g=1}^{G_{\text{unif}}}\mathcal{L}^i_{g,\text{unif}}\right], \quad (9)$$
$$\text{where} \quad \mathcal{L}^i_{g,\text{unif}} = \frac{1}{|p^i_g|}\sum_{t=1}^{|p^i_g|}\hat{A}^i_{g,\text{unif},t}.$$

For the exploratory group, $r^i_{g,t}$ scales updates based on the action likelihood under visual perturbations relative to the anchor:

$$J_{\text{rand}}(\theta) = \mathbb{E}_{\mathcal{D},\{p^i_g\}\sim\pi_\theta}\left[\frac{1}{G_{\text{rand}}}\sum_{g=1}^{G_{\text{rand}}}\mathcal{L}^i_{g,\text{rand}}\right], \quad (10)$$

$$\mathcal{L}^i_{g,\text{rand}} = \frac{1}{|p^i_g|}\sum_{t=1}^{|p^i_g|}\min\left(r^i_{g,t}\hat{A}^i_{g,\text{rand},t}, \bar{r}^i_{g,t}\hat{A}^i_{g,\text{rand},t}\right)$$
$$r^i_{g,t} = \frac{\pi_\theta(p^i_{g,t} \mid v^i_{\text{rand}}, q^i, p^i_{g,<t})}{\pi_\theta(p^i_{g,t} \mid v^i_{\text{unif}}, q^i, p^i_{g,<t})},$$
$$\bar{r}^i_{g,t} = \text{clip}(r^i_{g,t}, 1-\varepsilon, 1+\varepsilon).$$

Clipping $r^i_{g,t}$ establishes a **Visual Consistency Region (VCR)**. Unlike standard PPO (Schulman et al., 2017) which bounds temporal policy shifts, our formulation explicitly constrains the policy's *sensitivity to visual context changes*. Crucially, rather than suppressing the learning process, this VCR acts as a safeguard that *encourages bold exploration* of the reasoning manifold. It encourages that high-magnitude updates are safely executed only when the reasoning remains logically consistent across varied perspectives.

Final Objective. The complete objective combines both groups with a KL divergence regularization term to prevent excessive policy deviation:

$$J_{\text{final}}(\theta) = J_{\text{unif}}(\theta) + J_{\text{rand}}(\theta) - \alpha\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\theta_{\text{ref}}}), \quad (11)$$

where $J_{\text{unif}}(\theta)$ provides stable learning signals from consistent inputs, $J_{\text{rand}}(\theta)$ encourages exploration through diverse visual perspectives, and α controls the regularization.

Table 1: Comparisons on several video understanding benchmark datasets. Accuracy (%) is reported.

Models	Venue	Frames	Video Reasoning Benchmark			Video General Benchmark		
			VSI-Bench	VideoMMMU	MMVU (mc)	MVBench	TempCompass	VideoMME (w/o sub)
LLaMA-VID-8B (Li et al., 2024c)	ECCV'24	-	-	-	-	41.9	45.6	-
VideoLLaMA2-7B (Cheng et al., 2024)	arXiv'24	-	-	-	44.8	54.6	-	47.9
LongVA-7B (Zhang et al., 2024a)	TMLR'25	-	29.2	23.9	-	-	56.9	52.6
VILA-1.5-8B (Lin et al., 2024)	CVPR'24	-	28.9	20.8	-	-	58.8	-
Video-UTR-7B (Yu et al., 2025)	ICLR'25	-	-	-	-	58.8	59.7	52.6
LLaVA-OneVision-7B (Li et al., 2024a)	TMLR'24	-	32.4	33.8	49.2	56.7	-	58.2
Kangaroo-8B (Liu et al., 2024a)	arXiv'24	-	-	-	-	61.1	62.5	56.0
Qwen2.5-VL-7B-Instruct	arXiv'25	16	32.8	48.0	58.7	58.4	72.2	52.6
Qwen2.5-VL-7B w/ GRPO	arXiv'25	16	32.7	48.3	62.1	61.1	71.3	54.5
Video-R1-7B	NeurIPS'25	16	34.6	49.8	64.2	62.7	72.6	57.4
Ours-7B	-	16	**37.8**	**55.6**	**69.0**	**67.3**	**76.9**	**61.5**
Qwen2.5-VL-7B-Instruct	arXiv'25	32	30.1	48.1	60.0	59.0	72.6	56.6
Qwen2.5-VL-7B w/ GRPO	arXiv'25	32	30.5	49.5	63.4	62.6	73.3	58.0
Video-R1-7B	NeurIPS'25	32	35.8	52.3	63.8	63.9	73.2	59.3
Ours-7B	-	32	**39.3**	**55.8**	**69.6**	**67.6**	**77.4**	**62.9**

Table 2: Performance comparison across different backbones, methods, and frame counts. Accuracy (%) is reported.

Method	Qwen2.5-VL-7B-Instruct						MiMo-VL-7B-SFT						Qwen3-VL-8B-Instruct					
	MVBench		VideoMME		VideoMMMU		MVBench		VideoMME		VideoMMMU		MVBench		VideoMME		VideoMMMU	
	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f
Original	58.4	59.0	52.6	56.6	48.0	48.1	59.0	59.5	54.3	57.4	50.7	51.2	61.6	62.2	59.3	59.7	58.4	60.6
GRPO	61.1	62.6	54.5	58.0	48.3	49.5	60.7	60.0	56.5	57.8	51.5	52.2	62.4	63.0	59.9	60.8	60.0	61.2
Video-R1	62.7	63.9	57.4	59.3	49.8	52.3	61.5	61.1	56.6	57.3	52.4	53.1	63.3	63.7	61.8	62.1	60.7	61.5
Ours	67.3	67.6	61.5	62.9	55.6	55.8	65.4	65.9	60.8	61.8	55.9	56.5	68.6	69.0	64.4	65.3	64.3	64.8

4 Experiments

4.1 Experimental Configuration

Evaluation Benchmarks. Our assessment employs six video understanding benchmarks across two categories. For video reasoning benchmarks, we use VSI-Bench (Yang et al., 2025), VideoMMMU (Hu et al., 2025b), and MMVU (Zhao et al., 2025) (multiple-choice subset), which emphasize complex inferential tasks within video contexts. For general video understanding benchmarks, we evaluate on MVBench (Li et al., 2024b), TempCompass (Liu et al., 2024b), and VideoMME (Fu et al., 2025) (without subtitles), which assess both perceptual and reasoning capabilities. We conduct experiments across three backbone models: Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), MiMo-VL-7B-SFT (Xiaomi, 2025), and Qwen3-VL-8B-Instruct (Bai et al., 2025a), with varying frame counts (16 and 32 frames) to evaluate performance across different temporal resolutions.

Implementation and Training Configuration. Our experimental setup utilizes 8 NVIDIA H200 GPUs. Following standard practices (Shao et al., 2024), we set $G_{\text{unif}} = G_{\text{rand}} = 4$ ($G = 8$ per prompt) with a global batch size of 8. We employ a cosine learning rate schedule starting at 1×10^{-6} with a 10% warmup, *conducting all experiments across 3 independent seeds and reporting the averages*. To ensure numerical stability, rating parameters are set to $\mu = 1500$ and $\sigma = 500$. We train for 1,000 steps on the Video-R1-260K dataset (Feng et al., 2025), utilizing a context window of 4,000 tokens. During training, we use high-temperature ($T = 1.0$) sampling to ensure sufficient exploration; during inference, we employ near-deterministic decoding ($T = 0.01$) to evaluate reasoning stability, with all baselines re-evaluated under identical settings for a fair comparison. Hyperparameters $\varepsilon, \eta, \alpha$ are set to $0.3, 0.15, 0.04$ respectively. *Full implementation details are provided in the Appendix.*

4.2 Main Results

Baselines. We evaluate against two primary categories of baselines: (1) *Standard Video MLLMs*, including LLaMA-VID-8B (Li et al., 2024c), VideoLLaMA2-7B (Cheng et al., 2024), LongVA-7B (Zhang et al., 2024a), VILA-1.5-8B (Lin et al., 2024), Video-UTR-7B (Yu et al., 2025), LLaVA-OneVision-7B (Li et al., 2024a), and Kangaroo-8B (Liu et al., 2024a); (2) *RL-based Frameworks* for controlled cross-backbone comparisons, evaluating the pre-trained foundations (Original), standard GRPO (Shao et al., 2024), and the state-of-the-art video RL baseline Video-R1 (Feng et al., 2025).

Quantitative Results. As demonstrated in Tabs. 1 and 2, Our method consistently outperforms all es-



Question: What is the video telling when the burger placed in the upper right corner at the end of the video first appears? Options:
A. Beef with spices came from Russia to Germany. ✗
B. The steak began to be sandwiched between two pieces of bread. ✗
C. Steak burgers spread throughout the United States. ✓
D. The standardization of hamburgers. ✗

Video-R1: **<think>**Let me think about this for a moment. First, let's break down the options: ... Now, let's consider the video context. The video likely starts with the origin of the burger, ... Hmm, after this analysis, it seems that the most plausible context for the appearance of the burger in the upper right corner is that the steak began to be sandwiched between two pieces of bread. Let's verify one last time ... Therefore, the correct answer is B. **</think>**\n **<answer>B<\answer>**

Ours: **<think>**Let me think about this carefully. The video seems to be telling a story about the history of hamburgers ... The video starts with a drawing of a hamburger, and then it shows a map of the United States with various locations marked. This suggests that the video is tracing the spread of hamburgers across the country. Oh, I see. The video is likely telling us ... Let's verify this ... Therefore, the video is most likely telling us about the spreadout of hamburgers in the United States. **</think>**\n **<answer>C<\answer>**

Figure 4: Qualitative comparison on an instance of VideoMME (Fu et al., 2025). Given a video about hamburger history with multiple sequential frames, only our method correctly identifies the narrative at the specific timestamp (Option C), while Video-R1 selects incorrect option B.

Table 3: Ablation study on key components across different backbones and frame settings. We systematically evaluate the contribution of dual-group sampling (DG), entropy-rated ELO-based advantage (ELO-Adv), and cross-visual counterfactual weighting (CVW).

Method	Components			Qwen2.5-VL-7B-Instruct						Qwen3-VL-8B-Instruct					
	DG	ELO-Adv	CVW	MVBench		VideoMME		VideoMMMU		MVBench		VideoMME		VideoMMMU	
				16f	32f	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f
Baseline (GRPO)	×	×	×	61.1	62.6	54.5	58.0	48.3	49.5	62.4	63.0	59.9	60.8	60.0	61.2
+ DG	✓	×	×	63.5	64.9	57.1	60.9	50.7	52.2	64.6	65.9	62.0	62.5	61.6	62.3
+ ELO-Adv	×	✓	×	64.2	65.5	57.6	61.3	51.4	52.7	65.3	66.6	62.7	63.1	61.8	62.5
+ CVW	×	×	✓	63.2	64.7	56.8	60.4	50.5	51.9	64.1	65.5	62.3	62.0	61.3	62.2
+ DG + ELO-Adv	✓	✓	×	65.9	66.4	59.6	62.1	53.3	54.2	65.8	67.0	63.0	63.6	63.1	63.3
+ DG + CVW	✓	×	✓	64.8	65.9	58.3	61.6	52.4	52.7	64.7	66.4	62.5	63.0	62.5	63.0
+ ELO-Adv + CVW	×	✓	✓	66.0	66.3	59.9	62.0	53.5	54.4	65.4	66.8	63.1	63.4	63.0	63.5
Full Model (Ours)	✓	✓	✓	**67.3**	**67.6**	**61.5**	**62.9**	**55.6**	**55.8**	**68.6**	**69.0**	**64.4**	**65.3**	**64.3**	**64.8**

tablished baselines across six benchmarks. Notably, our framework surpasses standard GRPO by up to 7.3% and the state-of-the-art Video-R1-7B (Feng et al., 2025) by 4.1%–5.8% on key evaluation metrics. These substantial improvements hold across diverse backbone architectures, underscoring the generalizability and robustness of our dual-group exploration strategy.

Qualitative Results. The visualization in Fig. 4 highlights the superior temporal reasoning capabilities of Our method. In the challenging hamburger history scenario, the baseline Video-R1 (Feng et al., 2025) suffers from premature commitment, fixating on an isolated early event regarding bread sandwiching (Opt. B). Conversely, our method robustly contextualizes global spatial-temporal progressions across consecutive frames, precisely tracking the geographic dissemination of the narrative to deduce the correct judgment (Opt. C). This confirms that structured dual-group exploration effectively regularizes the multi-modal policy, enabling it to escape local optima and achieve high-fidelity visual consensus. *Additional quantitative and qualitative results are provided in the* ***Appendix***.

4.3 Analysis

Ablation Study. Tab. 3 evaluates the contribution of our three key components: dual-group sampling (DG), entropy-rated ELO-based advantage (ELO-Adv), and cross-visual counterfactual weighting (CVW). Each component independently improves the baseline GRPO, with ELO-Adv providing the strongest single-component gains (3.1% on MVBench, 3.1% on VideoMMMU). Pairwise combinations demonstrate synergistic effects, with ELO-Adv+CVW achieving 66.0% on MVBench. The full model integrating all three components delivers optimal performance: 67.3% on MVBench, 61.5% on VideoMME, and 55.6% on VideoMMMU (16f, Qwen2.5-VL), representing

6.2%, 7.0%, and 7.3% improvements over baseline GRPO. Consistent gains across both backbones and frame configurations demonstrate that each component plays a crucial role in establishing a robust visual consistency region and enhancing video reasoning.

Response Diversity and Exploration Quality. To quantify the link between exploration and performance, we analyze response diversity using Self-BLEU (Zhu et al., 2018b) and distinct-n (Li et al., 2016b) metrics in Tab. 4. Our full model consistently achieves the lowest Self-BLEU (0.33 on MVBench, 0.46 on VideoMME) and the highest distinct-n ratios. In contrast, the GRPO baseline exhibits severe *mode collapse* (Self-BLEU: 0.88 on VideoMME), where the model repeatedly generates near-identical outputs due to the lack of exploratory incentives. The correlation between increased diversity and the significant gains (7.0% over baseline) demonstrates that our framework successfully encourages the MLLM to navigate the visual-reasoning manifold more broadly, discovering more accurate reasoning pathways that are otherwise suppressed by standard RL protocols.

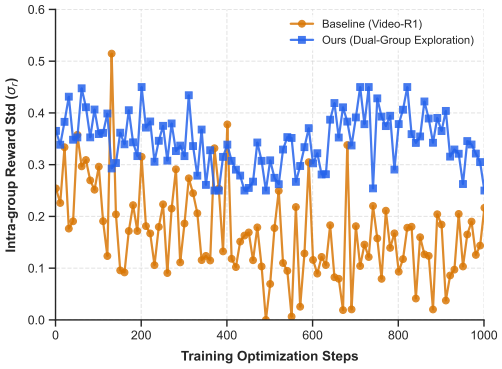


Figure 5: Real-time tracking of intra-group reward standard deviation (σ_r) across 1k optimization steps.

Empirical Verification of Advantage Vanishing. To verify the conceptualized "advantage vanishing" phenomenon, Fig. 5 tracks the real-time intra-group reward standard deviation (σ_r) during training. The baseline method, constrained by invariant video frames, suffers from recurrent mode collapse that forcefully drives $\sigma_r \to 0$ (e.g., at Steps 151 and 551). Analytically, given the GRPO standardization $A = (r - \mu)/(\sigma_r + \delta)$, this variance collapse degenerates the advantage into a uniform zero vector, completely neutralizing the policy gradient and halting learning. In contrast, Dual-Group Exploration breaks this visual determinism via stochastic temporal perturbations. By consistently maintaining a healthy advantage spread ($\sigma_r > 0.25$), our method preserves the critical directional gradients required for continuous policy improvement, thereby alleviating the distributional stagnation.

Counterfactual Weighting vs. Iterative Training. Tab. 5 shows multi-epoch training without adaptive regularization peaks early before visually overfitting (declining $60.1\% \to 59.9\%$ on VideoMME). Conversely, our single-epoch Cross-Visual Counterfactual Weighting (CVW) yields superior gains (+2.3% on VideoMMMU), confirming that establishing a robust visual consistency region safely outpaces brute-force iterative training.

Training Efficiency. Tab. 6 reveals our method requires only 8 responses per prompt. By bypassing Video-R1's costly 12-response requirement, our dual-group strategy matches GRPO's sampling budget with $< 1\%$ overhead. Consequently, it reduces total training time by 33.7% (Qwen2.5-VL) and 32.7% (Qwen3-VL) compared to Video-R1, achieving superior computational scalability. *See **Appendix** for additional experiments and analysis.*

Optimal Reference Anchoring We evaluate different standardization strategies in Tab. 7 to determine the most stable reference distribution. Reusing uniform group statistics for both groups outperforms separate standardization by up to 2.7% on VideoMMMU. This results from the uniform group acting as a *stable anchor*, ensuring that the rewards of the randomly sampled (exploratory) group are calibrated against a consistent baseline rather than shifting statistics that introduce instability.

Impact of Cross-Visual Counterfactual Weighting. Tab. 8 compares our Cross-Visual Counterfactual Weighting (CVW) against fixed weight baselines. Among fixed weights, 1.0 (equivalent to standard single-epoch GRPO lacking visual regularization) performs best at 65.9% on MVBench, 59.6% on VideoMME, and 53.3% on VideoMMMU, with deviations in either direction degrading performance. Our CVW mechanism substantially outperforms all fixed settings: +1.4% on MVBench (67.3%), +1.9% on VideoMME (61.5%), and +2.3% on VideoMMMU (55.6%). This validates that dynamically anchoring visual sensitivity through a Visual Consistency Region (VCR) provides the crucial regularization needed to empower bold, safe exploration, yielding superior reasoning performance over static constraints.

Impact of Frame Pool Expansion Factor. Tab. 9 investigates the impact of frame pool expansion

Table 4: Response diversity metrics across different methods on two benchmarks. Self-BLEU (Zhu et al., 2018a) measures similarity of varied answers (lower is better), while distinct-n (Li et al., 2016a) measures diversity of varied answers (higher is better).

Method	MVBench (16f)				VideoMME (16f)			
	Self-BLEU ↓	Distinct-1 ↑	Distinct-4 ↑	Accuracy ↑	Self-BLEU ↓	Distinct-1 ↑	Distinct-4 ↑	Accuracy ↑
Baseline (GRPO)	0.63	0.13	0.55	61.1%	0.88	0.09	0.27	54.5%
Ours w/o DG	0.54	0.20	0.62	66.0%	0.63	0.25	0.42	59.9%
Ours w/o ELO-Adv	0.42	0.26	0.69	64.8%	0.52	0.31	0.55	58.3%
Ours w/o CVW	0.45	0.19	0.64	65.9%	0.59	0.20	0.45	59.6%
Full Model (Ours)	**0.33**	**0.35**	**0.71**	**67.3%**	**0.46**	**0.44**	**0.59**	**61.5%**

Table 5: Impact of Cross-Visual Counterfactual Weighting and Training Epochs.

Mechanism	MVBench	VideoMME	VideoMMMU
Fixed Ratio (1 Epoch)	65.9	59.6	53.3
Fixed Ratio (3 Epochs)	66.2	60.1	53.7
Fixed Ratio (10 Epochs)	65.9	59.9	53.5
CVW (Ours, 1 Epoch)	**67.3**	**61.5**	**55.6**

Table 6: Comparison of training efficiency.

Method	Qwen2.5-VL-7B-Instruct		Qwen3-VL-8B-Instruct	
	# Responses	Training Time	# Responses	Training Time
GRPO	8	11.1h	8	14.2h
Video-R1	12	16.9h	12	21.4h
Ours	8	11.2h	8	14.4h

Table 7: Impact of different standardization strategies.

Standardization Strategy	MVBench	VideoMME	VideoMMMU
Separate Mean & Std for Each Group	65.5	60.1	52.9
Reuse Random Group Statistics	64.8	59.6	52.2
Cross-Group Statistics (Opposite)	63.0	58.3	50.4
Reuse Uniform Group Statistics (Ours)	**67.3**	**61.5**	**55.6**

Table 8: Comparison of cross-visual counterfactual weighting mechanisms.

Weight Mechanism	MVBench	VideoMME	VideoMMMU
Fixed weight = 0.6	64.8	59.2	52.4
Fixed weight = 0.8	65.3	59.2	52.9
Fixed weight = 1.0	65.9	59.6	53.3
Fixed weight = 1.2	65.7	59.5	52.7
Fixed weight = 1.4	64.9	59.0	52.1
CVW (Ours)	**67.3**	**61.5**	**55.6**

Table 9: Impact of frame pool expansion factor k on performance and computational cost. All experiments use 16 final frames.

Pool Size	MVBench	VideoMME	VideoMMMU
Fixed $k = 1$	63.6	58.9	52.0
Fixed $k = 2$	65.4	60.2	54.7
Fixed $k = 3$	65.6	59.7	54.5
Fixed $k = 4$	66.2	59.9	54.5
Random $k \in \{2, 3, 4\}$ (Ours)	**67.3**	**61.5**	**55.6**

factor k on model performance. When using a fixed expansion factor, we observe that larger pool sizes generally improve performance, with fixed $k = 4$ achieving 66.2% on MVBench compared to 63.6% for $k = 1$ (no expansion). However, the gains from simply increasing k show diminishing returns and inconsistent patterns across benchmarks, *e.g.*, VideoMME peaks at $k = 2$ (60.2%) and slightly declines for larger values, while VideoMMMU plateaus beyond $k = 2$ (54.7% vs 54.5% for $k = 3$ and $k = 4$). Remarkably, our strategy of randomly selecting $k \in \{2, 3, 4\}$ during training substantially outperforms any fixed k configuration, achieving 67.3%, 61.5%, and 55.6% on the three benchmarks respectively. This 1.1-1.3% improvement over the best fixed setting demonstrates that dynamically varying the pool size introduces beneficial training diversity that prevents overfitting to a specific sampling granularity. By exposing the model to different temporal sampling densities, our random k strategy enables more robust learning and better generalization across diverse video understanding tasks.

5 Conclusion

In this work, we mitigate critical exploration and regularization bottlenecks in video-based Multimodal LLMs. Existing paradigms suffer from *vanishing advantages*, and *visual overfitting* due to single-epoch training constraints. To address this, our ***Dual-Group Exploration*** RL framework revitalizes learning signals using dual-group sampling, an entropy-rated ELO advantage, and a novel cross-visual counterfactual weight. By establishing a *Visual Consistency Region* to anchor visual sensitivity, our method enables safe policy exploration. Achieving state-of-the-art performance across six benchmarks while reducing training time by 33%, we demonstrate that **consistency-guided exploration** is beneficial for MLLM video reasoning.

Limitations

While our method significantly reduces total training time by 33% and enhances response diversity, a few limitations remain. First, the dual-group sampling mechanism requires packing multiple complementary visual sequences within a single batch, which slightly increases the peak activation memory during the forward pass. Second, although our framework achieves a superior convergence rate in a single epoch, the overall computational footprint remains non-trivial when scaling to ultra-long videos or exceptionally extended reasoning chains. Further optimization regarding token pruning or episodic memory caches warrants future investigation to democratize resource-intensive multi-modal reinforcement learning.

Ethical Considerations

In alignment with the policy, we explicitly address the ethical implications, data integrity, and potential broader impacts of our work.

Data Integrity and Privacy. Our framework is developed and evaluated strictly using well-established, publicly available video-language open-source datasets (including VSI-Bench, MVBench, VideoMME, and VideoMMMU). We implemented a rigorous string-matching and n-gram based de-contamination protocol to eliminate overlapping instances between the training and evaluation splits, ensuring no data leakage. To the best of our knowledge, the source videos and text queries do not contain personally identifying information (PII), offensive content, or severe societal biases. We acknowledge and respect the original licensing and intellectual property rights of all curated benchmarks.

Hallucination Mitigations and Robustness. A key motivation of our Entropy-rated ELO-based advantage and Visual Consistency Region is to directly penalize stochastic, over-confident hallucinations that often plague multimodal alignment in real-world deployments. By explicitly enforcing a cross-visual consensus regularization, our method guides the multi-modal policy to evaluate evidence systematically before committing to a final assertion, thereby improving the truthfulness and safety of generative video-to-text systems.

Environmental Impact and Sustainability. Training large-scale Multimodal LLMs via reinforcement learning typically poses heavy carbon footprint and computational challenges. Our Dual-Group Exploration paradigm tackles this bottleneck by significantly reducing the total training duration (achieving a 33% runtime reduction compared to standard video RL baselines). This efficiency leap directly supports more sustainable, resource-friendly, and accessible AI model development, mitigating the computational barriers for smaller research institutions.

Dual-Use and Deployment Risks. As with any powerful generative Multimodal LLM technology, there remains a downstream risk of dual-use, such as the automated parsing of sensitive visual surveillance data or generating misaligned textual summaries under malicious prompts. We do not release any proprietary models or guarded data fractions, and we strongly advocate for deploying our exploration framework under strict safety guardrails and algorithmic auditing.

Reproducibility and Code Release

We are fully committed to open science and reproducibility. However, the codebase and models are currently undergoing a mandatory institutional clearance and security review process. Consequently, they cannot be included in the supplementary materials during the double-blind review period. We strictly commit to publicly releasing the complete code, models, and training scripts upon acceptance.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. 2025a. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. 2025b. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. 2024. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*.

Majid Ghasemi and Dariush Ebrahimi. 2024. Introduction to reinforcement learning. *arXiv preprint arXiv:2408.07712*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025a. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025b. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Wendi Li and Yixuan Li. 2024. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024c. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024a. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025. Capabilities of gpt-5 on multimodal medical reasoning. *artificial intelligence (AI)*, 10(11):12.

LLM-Core-Team Xiaomi. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, and Wenbing Tao. 2025. Unhackable temporal reward for scalable video MLLMs. In *The Thirteenth International Conference on Learning Representations*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yipeng Zhang, Yifan Liu, Zonghao Guo, Yidan Zhang, Xuesong Yang, Xiaoying Zhang, Chi Chen, Jun Song, Bo Zheng, Yuan Yao, and 1 others. 2024b. Llava-uhd v2: an mllm integrating high-resolution semantic pyramid via hierarchical window transformer. *arXiv preprint arXiv:2412.13871*.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018a. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018b. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. 2026. A.i.r.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *The Fourteenth International Conference on Learning Representations*.

A Extra Quantitative Results

Tab. 10 and Tab. 11 demonstrate the effectiveness of our proposed method across multiple video understanding benchmarks and backbone architectures. As shown in Tab. 10, our method (Ours-7B) achieves substantial improvements over existing approaches, outperforming the baseline Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) by 8.8% on MVBench (Li et al., 2024b), 4.0% on VideoMME (Fu et al., 2025), and 5.5% on VideoMMMU (Hu et al., 2025b). Notably, our approach surpasses both GRPO (Shao et al., 2024) and Video-R1 (Feng et al., 2025) methods across all benchmarks, with particularly significant gains on challenging reasoning tasks such as VSI-Bench (Yang et al., 2025) (+6.3% over Video-R1) and MMVU (Zhao et al., 2025) (+5.7% over Video-R1). Tab. 11 further validates the generalizability of our method across different backbone architectures. Consistent improvements are observed across all three backbones (Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), MiMo-VL-7B-SFT (Xiaomi, 2025), and Qwen3-VL-8B-Instruct (Bai et al., 2025a)), with our method achieving the highest performance in every configuration. For instance, on the Qwen3-VL-8B backbone, our method achieves 69.3% on MVBench and 66.6% on VideoMME, representing gains of 6.8% and 7.0% respectively over the original model. These results demonstrate that our approach provides robust and transferable improvements regardless of the underlying architecture, highlighting its effectiveness for video understanding tasks.

B Impact of Group Size

Tab. 12 examines the impact of group size G on performance, where each group is evenly divided into even and random subgroups. We observe a clear trend: increasing the group size from $G = 2$ to $G = 8$ yields substantial and consistent improvements across all benchmarks, with MVBench accuracy rising from 63.6% to 67.3%, VideoMME from 58.5% to 61.5%, and VideoMMMU from 53.3% to 55.6%. This demonstrates that larger groups provide more diverse response samples for advantage estimation, leading to more stable and effective learning signals. However, further increasing the group size to $G = 16$ and $G = 32$ produces only marginal gains (0.5% improvement on MVBench for $G = 16$) or even slight performance degradation (VideoMMMU drops from 55.9% to 55.2%). More importantly, larger group sizes incur significantly higher computational costs during training, as the number of forward passes scales linearly with G. Therefore, we select $G = 8$ as our default configuration, which achieves a favorable balance between performance and computational efficiency—it captures most of the benefits from diverse sampling while maintaining practical training costs for large-scale video understanding tasks.

C Impact of Temporal Order

Tab. 13 shows the importance of preserving temporal order in our dual-group sampling strategy. Disrupting temporal coherence, whether through unordered random sampling (64.7% on MVBench) or reversed-order sampling (63.3%), significantly degrades performance compared to the uniform sampling baseline (66.0%). In contrast, our temporally-ordered random sampling achieves superior results (67.3%), demonstrating that exploration through frame diversity must be balanced with temporal structure preservation. The consistent improvements across all three benchmark datasets (MVBench: +1.3%, VideoMME: +1.6%, VideoMMMU: +2.1%) confirm that video reasoning is tied to causal and temporal dynamics, so exploration must operate within the structural constraints of the video medium to be effective.

D Analysis of Entropy Weight

The sensitivity analysis reveals that the entropy weight parameter η exhibits a clear optimal range around 0.15, demonstrating robust yet nuanced effects across different video understanding benchmarks. Performance improves consistently as η increases from 0.0 to 0.15 across all three benchmarks, with MVBench showing a 2.5% gain (64.8 $\rightarrow$ 67.3), VideoMME a 3.2% improvement (58.3 $\rightarrow$ 61.5), and VideoMMMU a 3.2% increase (52.4 $\rightarrow$ 55.6), suggesting that incorporating entropy-based regularization meaningfully enhances model performance. However, the results also indicate diminishing returns and eventual performance degradation beyond the optimal point: while η values between 0.10 and 0.20 maintain near-peak performance with minimal variance ($\pm$0.3% on MVBench), further increases to 0.25 and 0.30 lead to consistent declines across all metrics. This inverted-U relationship suggests that moderate entropy weighting successfully balances the model's objectives, whereas excessive weighting ($\eta \geq$

Table 10: Comparisons among multiple MLLMs with 64 frames on several video understanding benchmark datasets. Accuracy (%) is reported.

Models	Venue	Frames	Video Reasoning Benchmark			Video General Benchmark		
			VSI-Bench	VideoMMMU	MMVU (mc)	MVBench	TempCompass	VideoMME (wo sub)
LLaMA-VID-8B (Li et al., 2024c)	ECCV'24	-	-	-	-	41.9	45.6	-
VideoLLaMA2-7B (Cheng et al., 2024)	arXiv'24	-	-	-	44.8	54.6	-	47.9
LongVA-7B (Zhang et al., 2024a)	TMLR'25	-	29.2	23.9	-	-	56.9	52.6
VILA-1.5-8B (Lin et al., 2024)	CVPR'24	-	28.9	20.8	-	-	58.8	-
Video-UTR-7B (Yu et al., 2025)	ICLR'25	-	-	-	-	58.8	59.7	52.6
LLaVA-OneVision-7B (Li et al., 2024a)	TMLR'24	-	32.4	33.8	49.2	56.7	-	58.2
Kangaroo-8B (Liu et al., 2024a)	arXiv'24	-	-	-	-	61.1	62.5	56.0
Qwen2.5-VL-7B-Instruct	arXiv'25	64	31.4	50.4	60.0	59.2	72.9	59.6
Qwen2.5-VL-7B w/ GRPO	arXiv'25	64	33.4	51.1	63.2	63.3	73.8	58.5
Video-R1-7B	NIPS'25	64	36.1	52.4	63.8	64.8	73.2	61.4
Ours-7B	-	64	**39.7**	**55.9**	**69.5**	**68.0**	**77.7**	**63.6**

Table 11: Performance comparison across different backbones and methods with 64 frames on video understanding benchmarks. Accuracy (%) is reported.

Method	Qwen2.5-VL-7B-Instruct			MiMo-VL-7B-SFT			Qwen3-VL-8B-Instruct		
	MVBench	VideoMME	VideoMMMU	MVBench	VideoMME	VideoMMMU	MVBench	VideoMME	VideoMMMU
Original	59.2	59.6	50.4	60.2	57.9	52.3	62.5	59.6	61.7
GRPO	63.3	58.5	51.1	60.8	58.8	53.5	63.3	61.4	62.6
Video-R1	64.8	61.4	52.4	62.5	58.3	54.0	64.2	63.0	62.9
Ours	**68.0**	**63.6**	**55.9**	**66.2**	**62.1**	**57.7**	**69.3**	**66.6**	**65.6**

Table 12: Impact of group size G on performance. Each group is split equally into even and random subgroups ($G_{\text{even}} = G_{\text{rand}} = G/2$).

Group Size	MVBench	VideoMME	VideoMMMU
$G = 2$	63.6	58.5	53.3
$G = 4$	66.6	60.2	55.0
$G = 8$ (Ours)	67.3	**61.5**	55.6
$G = 16$	**67.8**	61.4	**55.9**
$G = 32$	67.5	**61.5**	55.2

Table 13: Impact of temporal order in video frame sampling.

Sampling Strategy	MVBench	VideoMME	VideoMMMU
uniform Sampling	66.0	59.9	53.5
Random w/o Temporal Order	64.7	59.2	52.1
Random w/ Reversed Order	63.3	57.7	50.9
Random w/ Temporal Order (Ours)	**67.3**	**61.5**	**55.6**

Table 14: Sensitivity analysis of entropy weight η.

η	MVBench	VideoMME	VideoMMMU
0.0	64.8	58.3	52.4
0.05	66.6	60.7	55.0
0.1	67.0	**61.5**	55.4
0.15 (Ours)	**67.3**	**61.5**	**55.6**
0.20	67.2	61.2	55.4
0.25	67.0	61.0	55.2
0.30	66.7	61.0	55.0

Table 15: Sensitivity analysis of clipping bound ε.

ε	MVBench	VideoMME	VideoMMMU
0.15	66.7	60.6	54.3
0.20	67.0	60.9	55.0
0.25	**67.3**	61.2	**55.6**
0.30 (Ours)	**67.3**	**61.5**	**55.6**
0.35	67.2	**61.5**	55.4
0.40	67.0	61.0	55.2

0.20) likely over-constrains the model or introduces noise that interferes with primary task learning, validating the authors' choice of $\eta = 0.15$ as a well-calibrated hyperparameter.

E Analysis of Clipping Bound

The sensitivity analysis of the clipping bound parameter ε demonstrates that the model achieves optimal performance within a narrow range of $\varepsilon \in [0.25, 0.35]$, with performance exhibiting graceful degradation outside this interval. As ε increases from 0.15 to 0.30, all benchmarks show consistent improvement: MVBench gains 0.6% (66.7 $\rightarrow$ 67.3), VideoMME improves by 0.9% (60.6 $\rightarrow$ 61.5), and VideoMMMU increases by 1.3% (54.3 $\rightarrow$ 55.6), suggesting that moderate clipping bounds are crucial for stabilizing gradient updates without overly constraining the optimization process. The results reveal a performance plateau at peak values between $\varepsilon = 0.25$ and $\varepsilon = 0.35$, where metrics remain within $\pm 0.2\%$ across benchmarks, indicating robustness to small variations in the clipping threshold. However, both more conservative clipping ($\varepsilon = 0.15$) and more permissive

Table 16: Sensitivity analysis of the regularization parameter α.

α	MVBench	VideoMME	VideoMMMU
0	64.4	59.7	53.3
0.02	67.2	61.4	55.4
0.04 (Ours)	**67.3**	**61.5**	**55.6**
0.06	67.2	**61.5**	55.4
0.08	67.0	61.2	55.1
0.10	67.0	60.9	54.7

clipping ($\varepsilon = 0.40$) lead to performance degradation, with the former likely over-restricting gradient magnitudes and the latter potentially allowing destabilizing updates. This inverted-U relationship validates the authors' choice of $\varepsilon = 0.30$ as a well-tuned hyperparameter that balances training stability with sufficient gradient flow, particularly given its achievement of peak performance on VideoMME alongside competitive results on the other benchmarks.

The choice of $\varepsilon = 0.3$, while larger than the standard $0.1 \sim 0.2$, is empirically aligned with the exploratory variance introduced by our **cross-visual counterfactual weighting**. The results suggest that the Visual Consistency Region (VCR) provides sufficient regularization to maintain policy stability under broader clipping bounds. This configuration enables bold yet safe updates, directly contributing to the 33% faster convergence compared to traditional video RL baselines.

F Analysis of Regularization Parameter

The sensitivity analysis of the regularization parameter α reveals a sharp performance peak at $\alpha = 0.04$, with the model demonstrating high sensitivity to both under- and over-regularization. Performance improves dramatically as α increases from 0 to 0.04 across all benchmarks: MVBench gains 2.9% (64.4 $\rightarrow$ 67.3), VideoMME improves by 1.8% (59.7 $\rightarrow$ 61.5), and VideoMMMU increases by 2.3% (53.3 $\rightarrow$ 55.6), indicating that regularization is critical for effective model training and that the absence of regularization ($\alpha = 0$) leads to substantial performance degradation, likely due to overfitting or instability. The results show that performance remains relatively stable within a narrow range of $\alpha \in [0.02, 0.06]$, with metrics varying by less than 0.2% around the peak, suggesting some robustness to small deviations from the optimal value. However, increasing α beyond 0.06 causes consistent performance decline across all benchmarks, with $\alpha = 0.10$ resulting in losses of 0.3% on MVBench, 0.6% on VideoMME, and 0.9% on VideoMMMU relative to the optimal setting. This pattern indicates that excessive regularization progressively constrains the model's capacity to fit the training data, validating the authors' selection of $\alpha = 0.04$ as a carefully calibrated parameter that provides sufficient regularization to prevent overfitting while avoiding the underfitting that occurs with stronger regularization penalties.

G Chain-of-Thoughts Templates

Our prompt template (Fig. 6), building upon (Feng et al., 2025), structures the generation into explicit cognitive and output stages to elicit high-quality reasoning. By incorporating naturalistic internal dialogue (*e.g.*, "let's break it down"), the template encourages transparent Chain-of-Thought (CoT) reasoning, improving both interpretability and answer quality. The separation of reasoning (`<think>`) from final answers (`<answer>`) allows the model to mitigate complex temporal-spatial dependencies before commitment, providing a verifiable supervision signal during RL.

This unified template supports five distinct formats, like multiple choice, numerical, OCR, generation, and regression, enabling broad task coverage without architectural changes. To mitigate hallucinations and maintain factual grounding, the reasoning phase emphasizes self-reflection and verification. Crucially, the explicit monologue tags are utilized to decouple internal reasoning from the final output, enhancing logical faithfulness. To prevent spurious verbosity or "length-hacking," our RL rewards are strictly conditioned on final answer correctness, forcing the model to employ monologues for substantive reasoning rather than as a redundant strategy for maximizing rewards.

H Additional Details on Training Strategy

We employ Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), MiMo-VL-7B-SFT (Xiaomi, 2025), and Qwen3-VL-8B-Instruct (Bai et al., 2025a) as base multimodal large language models (MLLMs) for training. Following DeepSeek-R1 (Guo et al., 2025) and Video-R1 (Feng et al., 2025), we adopt a two-stage training paradigm: supervised fine-tuning (SFT) for cold-start initialization followed by reinforcement learning (RL) for reasoning refinement. Throughout both stages, we employ an image-video mixed training strategy to ensure ro-

{Question}

Approach this question with careful, deliberate reasoning as a human expert would. Engage in explicit internal dialogue using natural thought expressions such as ``let me think,'' ``wait,'' ``hmm,'' ``oh, I see,'' or ``let's break it down.'' Demonstrate your reasoning process transparently, including self-reflection, verification steps, and consideration of alternative approaches. Provide your complete reasoning chain within <think> and </think> tags, followed by your final answer within <answer> and </answer> tags.

{Output Template}

Multiple Choice: Provide only the single option letter (e.g., A, B, C, D) within <answer> </answer> tags.

Numerical: Provide the numerical value (e.g., 42 or 3.14) within <answer> </answer> tags.

OCR: Transcribe text from the image/video accurately and provide your transcription within <answer> </answer> tags.

Generation: Provide your complete text response within <answer> </answer> tags.

Regression: Provide the predicted numerical value (e.g., 42 or 3.14) within <answer> </answer> tags.

Figure 6: Prompt template for training and inference. The template in on the basis of Video-R1 (Feng et al., 2025) and instructs the model to engage in explicit chain-of-thought reasoning within `<think>` tags before providing task-specific answers in `<answer>` tags across five response formats: multiple choice, numerical, OCR, generation, and regression.

bust cross-modal generalization.

In the first stage, we perform SFT on the Video-R1-CoT-165k dataset (Feng et al., 2025) following established protocols (Feng et al., 2025; Chen et al., 2025b). This cold-start phase equips the model with foundational reasoning capabilities and familiarizes it with chain-of-thought (CoT) response templates across diverse modalities, establishing a baseline for subsequent reinforcement learning.

In the second stage, we further train the SFT-initialized models on the expanded Video-R1-260k dataset (Feng et al., 2025) using our proposed dual-group RL framework. This reinforcement learning phase is specifically designed to guide the model beyond the rigid pattern-matching behaviors induced by SFT, encouraging deeper exploration of the reasoning space and more flexible problem-solving strategies.

We define task-specific reward functions to evaluate model performance across different question types following (Zhan et al., 2025; Feng et al., 2025). Assume y is the ground truth annotation, and $\hat{y}$ is the prediction from the MLLM. For multiple choice questions, we use

$$R_{\text{mc}}(y, \hat{y}) = \mathbb{1}[y = \hat{y}] \tag{12}$$

where the reward is 1 if the predicted answer exactly matches the ground truth, and 0 otherwise. For numerical questions, the reward function is

$$R_{\text{num}}(y, \hat{y}) = \mathbb{1}[\lfloor n_y \rfloor_2 = \lfloor n_{\hat{y}} \rfloor_2] \cdot \mathbb{1}[d_y = d_{\hat{y}}] \tag{13}$$

where n_y denotes the numerical value extracted from string y, $\lfloor \cdot \rfloor_2$ rounds to 2 decimal places, and $d_y \in \{\text{True}, \text{False}\}$ indicates decimal presence in y. The reward is 1 only if both the rounded numerical values match and the decimal format is consistent. For OCR tasks, we define

$$R_{\text{ocr}}(y, \hat{y}) = \max\left(0, 1 - \frac{S + D + I}{N}\right) \tag{14}$$

where S, D, and I represent the minimum number of word-level substitutions, deletions, and insertions required to transform $\hat{y}$ into y, and N is the total number of words in the ground truth y. The reward decreases linearly with the edit distance. For free-form questions, the reward is

$$R_{\text{free}}(y, \hat{y}) = \frac{2 \cdot P \cdot R}{P + R} \tag{15}$$

where $P = \frac{|\text{LCS}(y,\hat{y})|}{|\hat{y}|}$ is the precision, $R = \frac{|\text{LCS}(y,\hat{y})|}{|y|}$ is the recall, and $\text{LCS}(y, \hat{y})$ denotes the

longest common subsequence between y and $\hat{y}$. This F1-score based on longest common subsequence measures semantic overlap. Finally, for regression tasks, we use

$$R_{\text{reg}}(y, \hat{y}) = 1 - \min\left(1, \frac{|n_{\hat{y}} - n_y| + \epsilon}{|n_y| + \epsilon}\right) \quad (16)$$

where n_y is the numerical value extracted from string y and $\epsilon = 10^{-9}$ prevents division by zero. The reward is based on relative error, with higher rewards for predictions closer to the ground truth value.

To enhance reasoning quality, we introduce a length-based reward mechanism that encourages sufficiently detailed responses. For each reasoning trajectory o^i from the i-th sample, if the predicted answer is correct and the response length exceeds a threshold l_{thres}, the model receives an additional reward $r_l = \omega$:

$$R^i = \begin{cases} R^i + \omega, & \text{if } o^i \text{ is correct and} \\ & \text{len}(o^i) \geq l_{\text{thres}} \\ R^i, & \text{otherwise} \end{cases}, \quad (17)$$

where we set $\omega = 0.3$ and $l_{\text{thres}} = 512$. This mechanism incentivizes the model to produce comprehensive reasoning chains without resorting to superficial shortcuts.

I Analysis of Reasoning Length and Computational Cost

Tab. 17 provides a comprehensive look at how length reward (LR) affects task performance and training duration. Crucially, the LR mechanism is designed to incentivize sufficient reasoning depth for complex queries rather than enforcing a rigid length requirement for every response. The observed performance gains as w increases to 0.3 suggest that providing an adequate "computational budget" allows the model to develop more thorough Chain-of-Thought trajectories, which are crucial for alleviating nuanced temporal dynamics. However, this does not result in uniform verbosity; for simpler tasks, the ELO-based consistency reward remains the dominant signal, ensuring that length is only extended when it contributes to reasoning accuracy. Furthermore, the training time data reveals a non-linear relationship between LR parameters and computational overhead. Despite increasing the threshold L_{max} from 320 to 768, the training time remains remarkably stable (fluctuating by only $\approx 2\%$), confirming that our Dual-Group framework successfully maintains efficiency by decoupling exploration quality from raw sequence length.

J Detailed Hyperparameters

To ensure the reproducibility of our results, we summarize the key training and architectural hyperparameters used in Dual-Group Exploration in Tab. 18.

To ensure a rigorous evaluation, we conducted a string-matching and n-gram based decontamination protocol. We verified that no overlapping pairs exist between the Video-R1-260K (Feng et al., 2025) training datasets and the evaluation benchmarks.

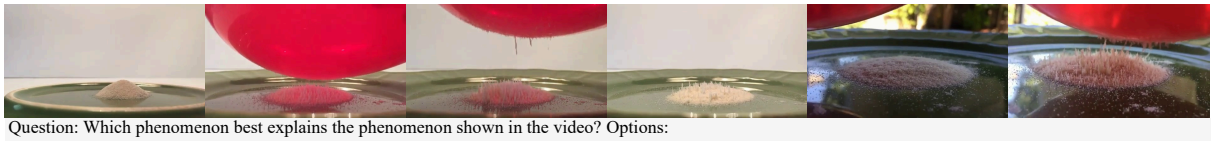
Calibration of Rating Parameters. The parameters $\mu = 1500$ and $\sigma = 500$ are calibrated to align with the standard ELO rating distribution. This ensures that the generated quality scores Q reside within a numerically stable range, facilitating consistent KL-divergence regularization. As demonstrated in our sensitivity analysis (see Appendix E), the Visual Consistency Region acts as an implicit stabilizer, making the framework robust to variations in these rating scales and preventing the performance degradation typically associated with heuristic-based reward shaping.

K Generalizability and Cross-Task Applicability

While our empirical evaluation focuses on video QA, the **Dual-Group Exploration** framework is *task-agnostic*. Its core components—the consistency-driven ELO reward and the trust-consistency policy update—are directly applicable to any video-to-text generation scenario. For instance, in **video summarization**, the ELO mechanism can penalize inconsistent or contradictory descriptions generated from different temporal samplings, thereby enforcing a more hallucination-free and coherent narrative. Similarly, for **temporal action localization**, the regression-based output format (supported by our CoT template) can be optimized via the same cross-visual consensus logic. This modular design ensures that our framework serves as a general-purpose paradigm for robust video RL, independent of specific task labels.

L Comparison with DAPO

To further demonstrate the architectural superiority of our method, we provide an extended compara-



Question: Which phenomenon best explains the phenomenon shown in the video? Options:
A. Magnetic attraction due to ferromagnetic properties of the particles. ✗
B. Thermal convection currents causing upward movement of particles. ✗
C. Electrostatic forces causing attraction between the charged balloon and the particles. ✓
D. Capillary action drawing the powder upward through liquid wicking. ✗
E. Chemical bonding between functional groups on the particles and balloon surfaces. ✗

Video-R1: **<think>**Let me think about this for a moment. The video shows a balloon being placed over a pile of pink particles, and the particles seem to be attracted to the balloon. This is a fascinating phenomenon, and I need to consider each option carefully.

First, let's break it down. The particles are not moving due to thermal convection currents (B), as there is no indication of heat or temperature changes. So, B is unlikely.

Next, let's consider electrostatic forces (C). While electrostatic forces can cause attraction, the video does not show any visible signs of charge buildup or discharge. The particles are not glowing or showing any signs of electrical activity, so C seems less likely.

Now, let's think about capillary action (D). Capillary action is the movement of a liquid through a narrow space due to the cohesive and adhesive properties of the liquid and the solid surfaces. However, there is no liquid present in the video, so D is not a viable option.

Finally, let's consider magnetic attraction (A). The particles appear to be ferromagnetic, meaning they can be attracted to the balloon due to their magnetic properties. The balloon, being a conductor, can induce a magnetic field that attracts the particles. This seems the most plausible explanation.

Hmm, I think I have it. The particles are likely attracted to the balloon due to their ferromagnetic properties. The balloon, being a conductor, can induce a magnetic field that attracts the particles. This aligns with the observed behavior in the video.

Oh, I see. The correct answer is A. The magnetic attraction due to ferromagnetic properties of the particles is the best explanation for the phenomenon shown in the video.

Let's verify one last time. The video does not show any signs of electrostatic forces, thermal convection, or capillary action, so A is the most logical choice.

After this deep reflection, I am confident that the correct answer is A. **</think>**\n **<answer>A<\answer>**

Ours: **<think>**Let me think about this for a moment. The video shows a balloon being brought close to a pile of pink powder, and the powder seems to be attracted to the balloon. This is a fascinating phenomenon, and I need to consider each option carefully.

A. Magnetic attraction due to ferromagnetic properties of the particles: This seems plausible because ferromagnetic materials can be attracted to magnets. However, the video doesn't show any magnets, and the balloon doesn't appear to be magnetic. So, this option seems unlikely.

B. Thermal convection currents causing upward movement of particles: This doesn't seem to fit the video. There's no indication of heat or temperature differences that would cause convection currents. The particles are moving upwards, but it doesn't look like it's due to heat.

C. Electrostatic forces causing attraction between the charged balloon and the particles: This is a strong possibility. When a balloon is rubbed against hair or clothing, it becomes charged and can attract small particles. The video shows the balloon being charged and then attracting the particles, which is consistent with electrostatic attraction.

D. Capillary action drawing the powder upward through liquid wicking: This doesn't seem to fit either. There's no liquid present in the video, and the powder is not being drawn upward through a liquid. The powder is moving upwards due to the balloon's charge.

E. Chemical bonding between functional groups on the particles and balloon surfaces: This is possible, but it's not as straightforward as electrostatic attraction. Chemical bonding would require a specific chemical interaction, which isn't evident in the video.

Hmm, after considering all the options, it seems that the most likely explanation is electrostatic attraction. The balloon becomes charged and attracts the particles due to the electrostatic forces. Oh, I see. The correct answer must be C. **</think>**\n **<answer>C<\answer>**

Figure 7: Qualitative comparison of reasoning processes between Video-R1 and our method on a example of MMVU. Video-R1 incorrectly concludes that magnetic attraction (Option A) explains the balloon-powder interaction, despite acknowledging the absence of magnets in the video and misidentifying the balloon as a conductor capable of inducing magnetic fields. In contrast, our method systematically evaluates each option, correctly identifies electrostatic forces (Option C) as the underlying mechanism by recognizing that rubbed balloons become charged and attract particles through electrostatic attraction, and explicitly eliminates magnetic attraction due to the absence of magnetic materials. This example demonstrates our method's improved scientific reasoning through better exploration.



Question: Which caption matches the video better?
Options:
A. The football is moving upwards. ✗
B. The football is moving from left to right in the video. ✓

Video-R1: **<think>**Let me think about this for a moment. The question is asking which caption matches the video better, and we have two options to choose from. First, let's consider option A: "The football is moving upwards." This option suggests that the football is being lifted or propelled into the air. In the image, we can see the football being kicked by a foot, and there is a clear indication of motion. The foot is in contact with the ball, and the ball appears to be in a position that suggests it is being kicked upwards. This aligns well with the idea that the football is moving upwards.

Now, let's consider option B: "The football is moving from left to right in the video." This option suggests a horizontal movement of the football. However, in the image, there is no clear indication of the football moving from left to right. The foot is positioned in a way that suggests a vertical kick, not a horizontal one. This option does not seem to match the visual evidence as well as option A does. Hmm, it seems that option A is the more accurate description of the video. The football is indeed moving upwards, which is a common action in soccer when a player kicks the ball. The image clearly shows the ball being lifted off the ground, which supports this conclusion.

Oh, I see. The correct answer should be the option that matches the visual evidence in the image. The football is moving upwards, and this is the most accurate description of the action in the video. Let's break it down one last time to verify. The foot is in contact with the ball, and the ball is in a position that suggests it is being kicked upwards. This aligns perfectly with option A, and it does not match the horizontal movement suggested by option B. After this deep reflection, I am confident that the correct answer is option A. Let me double-check one last time to ensure there are no mistakes. Yes, the football is moving upwards, and this is the most accurate description of the video.**</think>**\n **<answer>A<\answer>**

Ours: **<think>**Let me think about this for a moment. The video shows a close-up of a football being kicked by a player's foot. The foot is in the air, and the ball is also in the air, suggesting motion. The question asks which caption matches the video better, and we have two options to consider.

Option A states that the football is moving upwards. This could be true if the ball is being kicked with a force that propels it into the air, causing it to rise. However, the ball's trajectory is not entirely clear from the image alone. It could be moving upwards, but it could also be moving in a different direction.

Option B states that the football is moving from left to right in the video. This is a more specific description of the ball's movement. If we look at the image, the ball appears to be in a position that suggests it is moving from left to right. The foot is on the left side of the image, and the ball is on the right side, indicating a movement from left to right.

Hmm, let's break it down further. The ball is in the air, and the foot is kicking it. The direction of the kick is from left to right, which would cause the ball to move in that direction. The ball's position in the image supports this idea.

Oh, I see. The ball is indeed moving from left to right, as indicated by the position of the foot and the ball in the image. Therefore, the caption that matches the video better is the one that describes the ball's movement from left to right.

Let's verify this. The ball is not moving upwards in a way that is clearly visible in the image. The movement is more horizontal than vertical. Therefore, the best caption is the one that describes the ball's movement from left to right. **</think>**\n **<answer>B<\answer>**

Figure 8: Qualitative comparison on a temporal reasoning task requiring understanding of motion direction from TempCompass. Video-R1 incorrectly identifies the football as moving upwards (Option A), fixating on the vertical kick action while failing to track the ball's actual trajectory across the video frames. The model repeatedly reinforces its incorrect conclusion despite acknowledging ambiguity, stating "the foot is in contact with the ball, and the ball appears to be in a position that suggests it is being kicked upwards." In contrast, our method correctly identifies the horizontal left-to-right motion (Option B) by explicitly analyzing the spatial positions of the foot and ball across frames and reasoning about the directional implications: "The foot is on the left side of the image, and the ball is on the right side, indicating a movement from left to right." This example highlights our method's superior temporal reasoning capability.



Question: Which caption matches the video better? Options:
A. A person is picking up the pineapple. ✔
B. The pineapple is being pushed forward by a person. ✘

Video-R1: **<think>**Let me think about this for a moment. The question is asking which caption matches the video better, and we have two options to choose from. First, let's break it down and analyze the video content. In the video, we see a pineapple that has been cut in half, and a hand appears. The hand is interacting with the pineapple, but the exact nature of the interaction is not entirely clear from the still frames provided.

Now, let's consider the two options:

A. A person is picking up the pineapple.

- This option suggests that the person is lifting the pineapple off the surface. However, the video does not show the pineapple being lifted. The hand appears to be touching the pineapple, but it does not seem to be lifting it.

B. The pineapple is being pushed forward by a person.

- This option suggests that the pineapple is being moved forward by the person. The hand appears to be touching the pineapple and possibly applying some force, which could be interpreted as pushing it forward.

Hmm, it seems that the hand is more likely to be pushing the pineapple forward rather than lifting it. The video does not provide enough evidence to confirm that the pineapple is being lifted, but it does suggest that the hand is applying some force to move the pineapple.

Oh, I see. Based on the analysis, the more accurate description of the video is that the pineapple is being pushed forward by a person.

Let's verify this one last time. The hand is indeed touching the pineapple, and it appears to be applying some force, which aligns with the action of pushing it forward. After this reflection, I am confident that the better caption for the video is B.Therefore, the correct answer is B. Let's break it down one last time to ensure. The hand is touching the pineapple and appears to be applying force, which is consistent with pushing it forward. This aligns with option B.

Let's confirm one last time. The hand is touching the pineapple and applying force, which is consistent with pushing it forward. This confirms that the better caption is B. **</think>**\n **<answer>B<\answer>**

Ours: **<think>**Let me think about this for a moment. The video shows a pineapple that has been cut in half, revealing its bright yellow interior. The background is a vibrant pink, which contrasts with the yellow of the pineapple. A hand appears in the frame and interacts with the pineapple.

Now, let's consider the options given:

A. A person is picking up the pineapple.

B. The pineapple is being pushed forward by a person.

Hmm, if we look closely at the video, we can see that the hand is not just touching the pineapple but is actually lifting it. The hand is not pushing the pineapple forward; instead, it is lifting it off the surface. This action suggests that the person is picking up the pineapple.

Wait, let's break it down further. When we say "picking up," it implies that the object is being lifted from its current position. In this case, the hand is indeed lifting the pineapple from the surface. On the other hand, "pushing forward" would imply that the object is being moved in a forward direction, which is not what we see in the video. Oh, I see. The hand is not pushing the pineapple forward; it is lifting it. Therefore, the correct caption should reflect this action.

Let's verify this. The hand is clearly lifting the pineapple, not pushing it forward. This confirms that the correct caption is option

A. **</think>**\n **<answer>A<\answer>**

Figure 9: Qualitative comparison on a temporal action recognition task distinguishing between vertical and horizontal motion from TempCompass. Video-R1 incorrectly identifies the action as pushing the pineapple forward (Option B), reasoning that "the hand appears to be touching the pineapple and possibly applying some force, which could be interpreted as pushing it forward." Despite acknowledging uncertainty about whether the pineapple is being lifted, the model commits to the incorrect answer through repetitive self-reinforcement. In contrast, our method correctly identifies the action as picking up the pineapple (Option A) by carefully analyzing the motion trajectory: "if we look closely at the video, we can see that the hand is not just touching the pineapple but is actually lifting it." The model explicitly distinguishes between lifting (vertical motion) and pushing forward (horizontal motion), demonstrating superior spatial-temporal understanding. This example illustrates our method's ability to accurately interpret 3D motion dynamics from video sequences and resist premature conclusions when visual evidence requires careful scrutiny.



Question: What is the size of this room (in square meters)?
If multiple rooms are shown, estimate the size of the combined space.
Correct Answer: 38.7

Video-R1: **<think>**Let me think about this for a moment. The question is asking for the size of the room in square meters, and it mentions that there are multiple rooms shown. However, the image provided only shows one room, which is a living room.

Wait, I need to consider the possibility that the question is referring to the combined space of all the rooms shown in the image. The image shows a living room, a kitchen, and a hallway.

Hmm, let's break it down. The living room appears to be quite spacious, with a large TV, a sofa, and a dining table. The kitchen is also visible, with a stove, a sink, and a refrigerator. The hallway is narrow but connects the living room and the kitchen.

Oh, I see. To estimate the size, I need to consider the average dimensions of each room. The living room might be around 15 square meters, the kitchen around 10 square meters, and the hallway around 5 square meters.

Let's add these up: 15 + 10 + 5 = 30 square meters. This is a rough estimate, but it gives a good idea of the combined space.

Now, I need to verify if there are any other rooms shown in the image. After a careful look, I don't see any other rooms. So, the combined space of the rooms shown in the image is approximately 30 square meters.

Let me double-check my reasoning. The living room, kitchen, and hallway together make up the main areas shown in the image. There are no other rooms visible, so the estimate should be accurate.

Final estimate: The combined space of the rooms shown in the image is approximately 30 square meters. **</think>**\n **<answer>30.0<\answer>**

Ours: **<think>**Let me think about this for a moment. The image shows a living room, a kitchen, and a dining area. It's a bit tricky to estimate the exact size of each room, but I can make some educated guesses based on the visible features.

First, let's consider the living room. The room has a sofa, a coffee table, and a TV. The sofa looks like it could fit about 3 people comfortably, and the coffee table is a standard size. If we estimate the length of the room to be around 5 meters and the width to be about 4 meters, that would give us a rough estimate of the living room's area.

Next, let's look at the kitchen. The kitchen has a stove, a sink, and some cabinets. The stove is a standard size, and the cabinets look like they could be about 2 meters wide. If we estimate the length of the kitchen to be around 4 meters and the width to be about 3 meters, that would give us a rough estimate of the kitchen's area.

Finally, let's consider the dining area. The dining table looks like it could fit about 4 people, and the chairs are standard size. If we estimate the length of the dining area to be around 3 meters and the width to be about 2 meters, that would give us a rough estimate of the dining area's area.

Now, let's add up the areas. The living room is about 20 square meters, the kitchen is about 12 square meters, and the dining area is about 6 square meters. If we add these up, we get a total of about 38 square meters.

Hmm, that seems like a reasonable estimate. However, it's important to note that this is just an estimate based on the visible features. The actual size could be different. Oh, I see. The question asks for the size of the combined space, so we should add up the areas of all the rooms. The total combined space is about 38 square meters.

Let's break it down one more time to make sure. The living room is about 20 square meters, the kitchen is about 12 square meters, and the dining area is about 6 square meters. Adding these up, we get a total of about 38 square meters.

Yes, that seems correct. The combined space is about 38 square meters. **</think>**\n **<answer>38.0<\answer>**

Figure 10: Qualitative comparison on a spatial reasoning task from VSI-Bench requiring room dimension estimation. Video-R1 significantly underestimates the combined space at 30 square meters (ground truth: 38.7), with poor individual room estimates of 15, 10, and 5 square meters for the living room, kitchen, and hallway respectively. In contrast, our method achieves a more accurate estimate of 38 square meters by grounding dimensional judgments in concrete visual evidence: the living room at 20 square meters based on furniture dimensions (3-person sofa and coffee table suggesting 5m × 4m), the kitchen at 12 square meters from appliance sizes, and the dining area at 6 square meters from seating capacity. This example demonstrates that our enhanced exploration strategy enables more accurate extraction and integration of spatial information from video frames for quantitative spatial understanding tasks.

Table 17: Comprehensive Grid Search on Length Reward Parameters. We analyze the interaction between reward weight w and length reward threshold L_{thres}, showing their impact on performance and training efficiency.

Weight (w)	**Threshold (L_{thres})**			**Performance**			**Training Time**
	320	**512**	**768**	**MVBench**	**VideoMME**	**VideoMMMU**	
0.1	✓			66.4	60.2	54.8	**11.1h**
0.1		✓		66.8	60.4	54.8	**11.1h**
0.1			✓	66.8	60.5	55.0	11.2h
0.2	✓			66.9	60.9	55.2	**11.1h**
0.2		✓		66.9	61.0	55.3	11.2h
0.2			✓	67.1	61.2	55.3	11.3h
0.3	✓			67.2	61.1	53.4	**11.1h**
0.3 (Ours)		✓		**67.3**	**61.5**	**55.6**	11.2h
0.3			✓	**67.3**	61.4	**55.6**	11.4h

Table 18: Unified training and algorithmic configurations.

Hyperparameter	**Value**
Model & Training	
Global Batch Size	8
Optimization Steps	1000
Warm-up Steps	100
Learning Rate	1e-6 (Cosine)
Max Context Window	4000
Num Generations (G)	8 ($G_{\text{unif}} = 4, G_{\text{rand}} = 4$)
Training Temperature	1.0
Beta (β)	0.04
Entropy-rated ELO	
ELO Trade-off (η)	0.15
Rating Center (μ)	1500
Rating Spread (σ)	500
Length-based Bonus (ω)	0.3
Inference	
Inference Temperature	0.01
Inference Top-p	0.001
De-contamination	n-gram based

tive analysis against the recently proposed DAPO (Yu et al., 2026) baseline in Tab. 19. The empirical results reveal distinct behavioral trajectories: while DAPO delivers marginal yet consistent improvements over the vanilla GRPO benchmark by leveraging external alignment constraints, its capacity plateaus rapidly, yielding performance ceilings tightly clustered with Video-R1 (Feng et al., 2025). For instance, on the Qwen2.5-VL backbone with a 16-frame configuration, DAPO marginally edges out Video-R1 (+0.3% on MVBench) but trails or matches it on other temporal benchmarks.

In stark contrast, our Dual-Group Exploration framework consistently outperforms DAPO by a substantial margin across all evaluation matrices and frame counts. Notably, our method surpasses DAPO by **+4.3%** on MVBench (16f, Qwen2.5-VL) and achieves a pronounced gain of **+5.1%** on the scale-expanded Qwen3-VL backbone. This cross-backbone performance gap strongly indicates that simply executing policy alignment without regularizing the visual input layer introduces immediate optimization stagnation under complex temporal data. Conversely, by effectively restructuring the exploration manifold through active cross-visual consensus and visual consistency regions, our framework unlocks generalized reasoning capabilities that are unachievable by pure output-level data optimization methods.

M Disentangling Visual Consensus from Frame Exposure

A potential confounding hypothesis is that our performance scaling originates merely from an increased exposure to visual tokens rather than the intrinsic scaling of the policy optimization boundary. To decouple these factors, we conduct a rigorous stress-test in Tab. 20 by scaling the input density of fixed-frame baseline methods up to quadruple granularity ((16×4)f or 64 frames) and matching them against our method restricted to a baseline 16f budget.

The empirical scaling curves yield two pivotal insights. First, for standard text-centric extensions like GRPO and Video-R1 (Feng et al., 2025), simply scaling the raw visual context injection exhibits

Table 19: Performance comparison including DAPO. Accuracy (%) is reported across different backbones, frame counts, and methods.

Method	Qwen2.5-VL-7B-Instruct						Qwen3-VL-8B-Instruct					
	MVBench		VideoMME		VideoMMMU		MVBench		VideoMME		VideoMMMU	
	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f	16f	32f
GRPO	61.1	62.6	54.5	58.0	48.3	49.5	62.4	63.0	59.9	60.8	60.0	61.2
DAPO	63.0	63.5	56.6	58.9	49.4	52.4	63.5	63.9	60.3	61.1	60.8	61.2
Video-R1	62.7	63.9	57.4	59.3	49.8	52.3	63.3	63.7	61.8	62.1	60.7	61.5
Ours	**67.3**	**67.6**	**61.5**	**62.9**	**55.6**	**55.8**	**68.6**	**69.0**	**64.4**	**65.3**	**64.3**	**64.8**

Table 20: Quantitative comparison of Reasoning Quality vs. Extra Frame Exposure using the Qwen2.5-VL-7B backbone. Accuracy (%) is reported across key video understanding benchmarks.

Method	Frame Strategy	Effective Frames	MVBench	VideoMME	VideoMMMU	Avg. (Δ_{base})
GRPO	Fix Frame	16f	61.1	54.5	48.3	54.6 ($\pm$0.0)
GRPO	Fix Frame	(16 $\times$ 2)f	62.6	58.0	49.5	56.7 (+2.1)
GRPO	Fix Frame	(16 $\times$ 4)f	63.3	58.5	51.1	57.6 (+3.0)
Video-R1	Fix Frame	16f	62.7	57.4	49.8	56.6 ($\pm$0.0)
Video-R1	Fix Frame	(16 $\times$ 2)f	63.9	59.3	52.3	58.5 (+1.9)
Video-R1	Fix Frame	(16 $\times$ 4)f	64.8	61.4	52.4	59.5 (+2.9)
Ours	**Dynamic Frame**	**16f**	**67.3**	**61.5**	**55.6**	**61.5 (+4.9)**

severe law of diminishing returns. Quadrupling the temporal frame tokens only provides a marginal average accuracy boost of **+3.0%** for GRPO and **+2.9%** for Video-R1, highlighting that unregularized policy learning suffers from severe information saturation.

Second, and most phenomenally, our framework operating at the baseline **16f** budget substantially outperforms the heavily over-loaded **64f** Video-R1 baseline on average (**61.5%** vs. **59.5%**). While the 64-frame Video-R1 baseline closely approaches our accuracy on specific local benchmarks like VideoMME due to raw dense frame coverage (61.4% vs. 61.5%), it introduces a catastrophic $4\times$ computation cost and attention KV-cache inflation during the forward-backward pass. In contrast, our method establishes a robust visual consistency region that enables more accurate extraction and integration of sparse spatial-temporal information without relying on expensive information exposure. This underscores that maximizing the exploration quality per token is drastically more mathematically sound and computationally sustainable than brute-force context expansion.

N Extra Qualitative Visualizations

Fig. 7, Fig. 8, Fig. 9, and Fig. 10 present qualitative comparisons illustrating fundamental differences in reasoning quality between Video-R1 and our method across diverse video Q&A datasets such as MMVU (Zhao et al., 2025), TempCompass (Liu et al., 2024b), VSI-Bench (Yang et al., 2025). A recurring pattern emerges where Video-R1 exhibits premature commitment to incorrect answers through repetitive self-reinforcement, often fixating on superficial visual cues while failing to systematically eliminate alternatives or accurately extract quantitative information from visual evidence. In Fig. 7, Video-R1 demonstrates flawed scientific reasoning by conflating magnetic and electrostatic phenomena, incorrectly attributing ferromagnetic properties to a balloon despite explicitly acknowledging the absence of magnets, representing a clear failure in logical consistency. Fig. 8 and Fig. 9 reveal Video-R1's weakness in temporal reasoning: the model struggles to distinguish between instantaneous actions and sustained motion trajectories (Fig. 8) and conflates horizontal versus vertical motion patterns (Fig. 9), suggesting inadequate integration of spatial-temporal information across video frames. Fig. 10 further exposes Video-R1's limitations in spatial reasoning, where the model

significantly underestimates room dimensions (30 vs. 38.7 square meters ground truth), failing to ground dimensional estimates in concrete visual evidence such as furniture sizes and spatial proportions. In contrast, our method demonstrates more robust reasoning across all tasks by systematically evaluating all options before reaching conclusions, explicitly grounding interpretations in observable evidence rather than assumptions, accurately tracking motion dynamics across temporal sequences, extracting quantitative spatial information from visual cues, and maintaining logical consistency throughout the reasoning process. These qualitative improvements align with our quantitative results and demonstrate that our dual-group reinforcement learning framework with entropy regularization successfully encourages deeper exploration of the solution space, enabling the model to escape local optima and discover more accurate reasoning pathways. The examples collectively highlight that effective video reasoning requires genuine understanding of physical principles, temporal causality, spatial-temporal relationships, and quantitative spatial reasoning beyond simple pattern recognition, capabilities that emerge from our method's enhanced exploration strategy.