VIDEO-KTR: REINFORCING VIDEO REASONING VIA KEY TOKEN ATTRIBUTION

Ziyue Wang[1,2], Sheng Jin[1], Zhongrong Zuo[1], Jiawei Wu[3], Han Qiu[1], Qi She[1], Hao Zhang[4,†], Xudong Jiang[2,†]

[1]ByteDance
[2]School of Electrical and Electronic Engineering, Nanyang Technological University
[3]National University of Singapore
[4]College of Computing and Data Science, Nanyang Technological University

ziyue005@e.ntu.edu.sg, hzhang26@outlook.com, exdjiang@ntu.edu.sg

ABSTRACT

Reinforcement learning (RL) has shown strong potential for enhancing reasoning in multimodal large language models, yet existing video reasoning methods often rely on coarse sequence-level rewards or single-factor token selection, neglecting fine-grained links among visual inputs, temporal dynamics, and linguistic outputs, limiting both accuracy and interpretability. We propose **Video-KTR**, a modality-aware policy shaping framework that performs selective, token-level RL by combining three attribution signals: (1) *visual-aware* tokens identified via counterfactual masking to reveal perceptual dependence; (2) *temporal-aware* tokens detected through frame shuffling to expose temporal sensitivity; and (3) *high-entropy* tokens signaling predictive uncertainty. By reinforcing only these key tokens, Video-KTR focuses learning on semantically informative, modality-sensitive content while filtering out low-value tokens. Across five challenging benchmarks, Video-KTR achieves state-of-the-art or highly competitive results—42.7% on Video-Holmes (surpassing GPT-4o)—with consistent gains on both reasoning and general video understanding tasks. Ablation studies verify the complementary roles of the attribution signals and the robustness of targeted token-level updates. Overall, Video-KTR improves accuracy and interpretability, offering a simple, drop-in extension to RL for complex video reasoning. Our code and models are available at https://github.com/zywang0104/Video-KTR.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful paradigm for enhancing long-chain reasoning in large language models (LLMs) (OpenAI, 2024; DeepSeek-AI, 2025; Yang et al., 2025). In particular, GRPO (Shao et al., 2024) exhibits strong performance on complex reasoning tasks. Building on this success, RL has been extended to multimodal LLMs (MLLMs), achieving notable gains on visual understanding through sample diversification (Wang et al., 2025a; Leng et al., 2025), tailored reward design (Tan et al., 2025; Shen et al., 2025), and advanced optimization strategies (Deng et al., 2025; Zhang et al., 2025; Dang et al., 2025). Yet, these advances primarily target static images, leaving video reasoning comparatively underexplored. Recent work mitigates this gap by integrating temporal supervision, *e.g.*, contrasting

[†] Correspondence to Hao Zhang and Xudong Jiang.

predictions on ordered and shuffled frames, to strengthen temporal sensitivity (Feng et al., 2025), and by embedding spatio-temporal priors into reward mechanisms to better align spatial grounding with temporal ordering (Li et al., 2025b; Sun et al., 2025). Other studies build RL-driven tool agents capable of decomposing and solving long-form video queries (Tian et al., 2025), as well as scalable reward designs that adapt to video complexity via difficulty-aware or temporally robust objectives (Park et al., 2025; Chen et al., 2025b).

However, video reasoning still struggles to accurately model temporal dynamics and exploit visual cues. Many approaches overlook explicit temporal dependencies and causal structures—essential for effective reasoning (Tian et al., 2025; Huang et al., 2025). Some methods (Feng et al., 2025) introduce temporal-specific constraints, such as penalizing predictions on shuffled frames, yet these rely on coarse global assumptions. This design ignores tasks solvable by static cues, introducing optimization noise that may hinder training. Moreover, many approaches fail to ensure fine-grained semantic alignment between visual inputs and output tokens. Without explicit modeling of modality-specific dependencies, models underutilize visual evidence and over-rely on linguistic priors (Wang et al., 2025d; Li et al., 2025b), increasing hallucination risks. Even token-level RL prioritizing high-uncertainty tokens via entropy (Wang et al., 2025b) often lack modality awareness, limiting their ability to capture visual–temporal dependencies. The absence of precise token-level correspondence across modalities not only reduces reasoning accuracy but also limits interpretability.

To address these challenges, we propose **Video-KTR**, a modality-aware policy shaping framework for token-level optimization in video reasoning. Unlike prior methods that impose coarse, global temporal constraints (Feng et al., 2025; Dang et al., 2025) or rely solely on entropy-based token selection (Wang et al., 2025b), Video-KTR explicitly models token sensitivity to visual and temporal perturbations, updating only critical tokens. We categorize these tokens as: (1) *visual-aware tokens*, identified via counterfactual masking to capture reliance on perceptual input; (2) *temporal-aware tokens*, detected through frame shuffling to reflect sensitivity to temporal order and causality; and (3) *high-entropy tokens*, characterized by predictive uncertainty, complementing the first two categories. This selective updating focuses learning capacity on the most informative reasoning steps, reducing interference from redundant tokens. As a result, Video-KTR significantly enhances both accuracy and integrated multimodal reasoning in complex video tasks.

To assess Video-KTR, we conduct extensive experiments on representative video reasoning benchmarks. Results demonstrate SOTA or highly competitive performance across datasets. On Video-Holmes, our method achieves an overall accuracy of 42.7%, surpassing GPT-4o (42.0%). Moreover, ablation experiments show that modality-aware token attribution, selective reinforcement learning, and targeted perturbation jointly enhance the exploitation of visual and temporal cues. These findings confirm that Video-KTR's fine-grained, token-level optimization provides an effective and interpretable approach to advancing video reasoning.

Our contributions are twofold: 1) We employ counterfactual analysis to unveil modality-sensitive critical tokens in video reasoning, and propose Video-KTR tailored for token-level optimization in video reasoning tasks. To the best of our knowledge, this is the first work to integrate modality-aware token selection into RL for video reasoning. 2) Video-KTR achieves SOTA performance on multiple video reasoning benchmarks, including 42.7% on Video-Holmes dataset—matching GPT-4o and significantly outperforming existing open-source baselines. Ablation studies confirm the effectiveness of our modality-attribution signals, token selection strategy, and framework design, alongside the robustness and generalizability of Video-KTR.

2 Related Work

Reinforcement learning has emerged as a powerful paradigm for enhancing the reasoning abilities of LLMs (OpenAI, 2024; DeepSeek-AI, 2025; Yang et al., 2025), while GRPO (Shao et al., 2024) and its variants are particularly influential in optimizing outcome-level correctness. Upon these advances, RL has been increasingly extended to multimodal LLMs (Wang et al., 2025c; Huang et al., 2025; Leng et al., 2025; Meng et al., 2025; Yuan et al., 2025; Zhang et al., 2025; Chen et al., 2025a). However, these methods re-

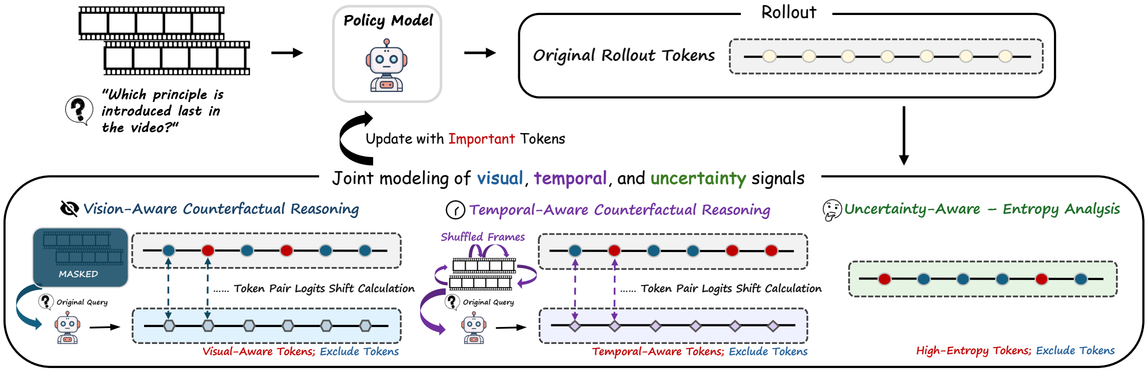

Policy Model
Rollout
Original Rollout Tokens
"Which principle is introduced last in the video?"
Update with Important Tokens
Joint modeling of visual, temporal, and uncertainty signals
Vision-Aware Counterfactual Reasoning
MASKED
Original Query
...... Token Pair Logits Shift Calculation
Visual-Aware Tokens; Exclude Tokens
Temporal-Aware Counterfactual Reasoning
Shuffled Frames
Original Query
...... Token Pair Logits Shift Calculation
Temporal-Aware Tokens; Exclude Tokens
Uncertainty-Aware – Entropy Analysis
High-Entropy Tokens; Exclude Tokens

Figure 1: Overview of the Video-KTR framework. The model identifies key tokens based on entropy, visual, and temporal signals, and updates only those tokens during reinforcement learning.

main centered on visual perception, with limited attention to temporal dynamics. Recent research proposes specialized RL strategies tailored for video reasoning, including temporally-aware fine-tuning by contrasting predictions from shuffled versus ordered frames (Feng et al., 2025), enhancing spatio-temporal reasoning via structured rewards (Li et al., 2025b; Sun et al., 2025), introducing tool-based agent frameworks for handling long-form video tasks (Tian et al., 2025), and developing scalable reward designs such as difficulty-aware regression and robust temporal reward formulations (Park et al., 2025; Chen et al., 2025b). Despite notable progress, these methods predominantly rely on coarse sequence-level rewards or limited token-level selection methods without explicit modeling of modality-specific dependencies. In contrast, we propose to jointly model complementary visual, temporal, and high-entropy token attribution signals, enabling more precise token-level credit assignment and improving reasoning performance. To the best of our knowledge, this is the first work to unify multiple token attribution strategies for modality-sensitive RL in Video-MLLMs.

3 Methodology

Recent RL work for Video-MLLMs (Feng et al., 2025; Chen et al., 2025b) mainly relies on coarse sequence-level rewards or simple token-level heuristics that capture only predictive uncertainty, without modeling video-specific structures like visual–temporal coherence or causal dependencies. In this work, we propose **Video-KTR**, a modality-aware policy shaping framework (Figure 1) that performs fine-grained RL through token-level attribution. We use counterfactual analysis to identify tokens conditioned on visual cues and temporal order, and exploit predictive entropy to detect reasoning-critical tokens. By selectively updating these tokens, Video-KTR amplifies reasoning-sensitive signals, improving both accuracy and interpretability.

3.1 Multi-Perspective Token Importance Analysis

A central challenge in video reasoning is identifying which outputs truly rely on visual or temporal cues. Prior attribution methods, such as entropy-based approaches (Wang et al., 2025b), expose uncertain tokens but cannot distinguish modality-specific dependence from generic reasoning difficulty. To address this, we propose a modality-aware attribution framework that jointly captures visual dependence, temporal sensitivity, and predictive uncertainty, enabling precise identification of key reasoning tokens.

Visual-Aware Tokens. Tokens closely associated with visual input are vital for multimodal reasoning, as they align language outputs with perceptual evidence and enable key functions like reference resolution, object recognition, and appearance description. These capabilities are particularly important in tasks that involve spatial localization, appearance-based inference, and dynamic scene understanding. Effective targeted learning thus hinges on accurately identifying output components that genuinely rely on visual information.

To identify these visual-aware tokens, we introduce a visual attribution mechanism based on counterfactual perturbation, which quantifies each token's sensitivity to visual input by masking the video and measuring the resulting shift in logits. Given a video input $\mathbf{v}$ and textual prompt $\mathbf{t}$, the model generates a response sequence $\hat{y} = (y_1, y_2, \ldots, y_T)$, with decoder logits z_i for each token y_i. We first compute logits under both full-context and masked-visual settings, extract the target token's logit at each position as:

$$\mathbf{z_i^{\text{visible}}} = f_\theta(y_i \mid y_{<i}, \mathbf{v}, \mathbf{t}), \quad \mathbf{z_i^{\text{masked}}} = f_\theta(y_i \mid y_{<i}, \tilde{\mathbf{v}}, \mathbf{t}),$$

where f_θ is the decoder network, $\mathbf{v}$ denotes the original visual input and $\tilde{\mathbf{v}} = \mathbf{0}$ indicates visual features are masked. Then, the visual attribution score is calculated by the log-probability shift as:

$$\Delta_i^{\text{vis}} = \left| \log \operatorname{softmax}(\mathbf{z_i^{\text{full}}})_{y_i} - \log \operatorname{softmax}(\mathbf{z_i^{\text{masked}}})_{y_i} \right|,$$

where Δ_i^{vis} assesses the visual dependence of i-th token. A larger Δ_i^{vis} indicates that the prediction of token y_i is strongly dependent on visual input. Depicted in Figure 2a, tokens such as "person," "door," and "blue" exhibit high visual sensitivity and are prioritized during training to strengthen perceptual grounding.

Temporal-Aware Tokens. Tokens sensitive to the temporal structure of videos are crucial for understanding event order, causality, and temporal dynamics. They allow the model to distinguish temporal relations, reason about progression, and track changes over time. To enhance temporal reasoning, it is essential to identify output tokens significantly affected by the sequence of visual events. Unlike T-GRPO (Feng et al., 2025), which applies sequence-level penalties to temporally shuffled inputs, we instead compute logit shifts induced by frame shuffling to locate tokens most responsive to temporal cues, enabling more fine-grained and targeted model optimization. To identify temporal-aware tokens, we disrupt the input's temporal order by shuffling the video frames, thereby breaking the original event sequence, and compute the resulting token-wise logit shifts. Given a video sequence $\mathbf{v} = (v_1, v_2, \ldots, v_F)$ and a textual prompt $\mathbf{t}$, the model generates a response sequence $\hat{y} = (y_1, y_2, \ldots, y_T)$, with decoder logits z_i for each token y_i. Similarly, we extract the target token's logit at each position as:

$$\mathbf{z_i^{\text{ordered}}} = f_\theta(y_i \mid y_{<i}, \mathbf{v}, \mathbf{t}), \quad \mathbf{z_i^{\text{shuffled}}} = f_\theta(y_i \mid y_{<i}, \hat{\mathbf{v}}, \mathbf{t}),$$

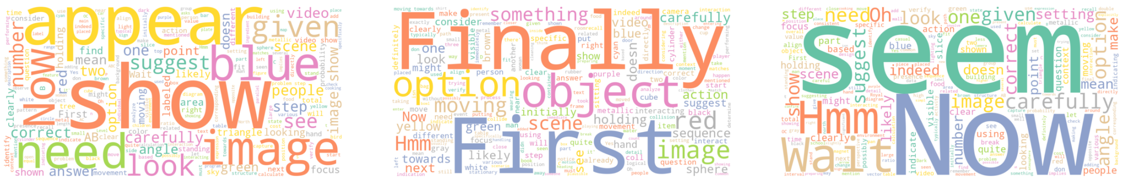
when f_θ denotes the decoder network, and the temporal attribution score for each token is computed as:

$$\Delta_i^{\text{temp}} = \left| \log \operatorname{softmax}(\mathbf{z_i^{\text{ordered}}})_{y_i} - \log \operatorname{softmax}(\mathbf{z_i^{\text{shuffled}}})_{y_i} \right|.$$

Similar to visual-aware tokens, a larger Δ_i^{temp} indicates stronger reliance on the video's temporal structure. Tokens such as "first," "then," and "appear" with high Δ_i^{temp} (Figure 2b) are highly sensitive to event progression and transitions, making them essential for modeling event flow and long-range dependencies beyond isolated frames. This attribution mechanism not only pinpoints temporally sensitive reasoning cues but also highlights where temporal understanding is most critical. In video question answering, many queries depend on capturing action sequences and event relations. For instance, answering "*What did the person do after entering the room?*" requires precise temporal reasoning. Our temporal-aware attribution explicitly reveals such dependencies, guiding reinforcement learning toward causally significant moments.

Entropy-Aware Tokens. While visual and temporal attributions capture modality-specific dependencies, they may miss reasoning-critical tokens unrelated to perception or sequence. Thus, we incorporate predictive entropy (Wang et al., 2025b) as a logic-aware criterion to identify low-confidence tokens:

$$\mathcal{H}(i) = -\sum_{v} p(z_i = w) \log p(z_i = w),$$



(a) Visual-aware tokens.

(b) Temporal-aware tokens.

(c) Entropy-aware tokens.

Figure 2: An illustration of critical tokens identified by each attribution strategy.

where w denotes token index. High-entropy tokens—*e.g.*, "however," "wait"—often mark discourse pivots or decision points critical for reasoning. Illustrated in Figure 2c, they capture semantic ambiguity, conflicting cues, or implicit causality, and offer a logic-aware means of identifying reasoning signals beyond visual and temporal cues. We incorporate it into token-level credit assignment as an empirically validated strategy.

3.2 Token Selection and Policy Update

Building on the token importance estimation, we explore how to integrate this information into the RL process effectively. The central challenge lies in *how to leverage the identified key tokens to enhance the training efficiency*. Conventional RL frameworks like GRPO treat all tokens equally during reward assignment, which dilutes learning signals—particularly in long video reasoning sequences—and thus hampers training efficiency. We introduce a **token-based policy shaping** mechanism that selectively reinforces semantically critical tokens during training. To be specific, we select the top $r\%$ of tokens from each attribution strategy and take their union, $\mathcal{S} = \mathcal{S}_{\text{vis}} \cup \mathcal{S}_{\text{temp}} \cup \mathcal{S}_{\text{ent}}$, as the set of **key reasoning tokens**. These tokens collectively capture the most essential semantic cues for visual grounding, temporal understanding, and decision uncertainty, thereby enabling more focused and efficient policy updates.

We adopt the GRPO framework (Shao et al., 2024), which improves the current policy π_θ by contrasting its outputs with a set of reference completions sampled from the previous policy $\pi_{\theta_{\text{old}}}$. Given a batch of G questions $q \sim P(Q)$ and corresponding rollouts $o_i{}_{i=1}^{G} \sim \pi\theta_{\text{old}}(\cdot|q)$, the GRPO objective is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}\cdot\hat{A}_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\cdot\hat{A}_{i,t}\right)\right], \tag{1}$$

where the KL divergence term is omitted for simplicity, $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ is the token-level likelihood ratio, and $\hat{A}_{i,t}$ denotes the advantage estimate computed via the normalized group reward for all tokens in o_i. To prioritize semantically important reasoning tokens during learning, we apply a binary mask $m_{i,t} \in \{0, 1\}$ to selectively update tokens. This mask is constructed from the union of top-ranked tokens identified by visual-aware, temporal-aware, and entropy-aware attributions. The resulting objective is:

$$\mathcal{J}_{\text{Video-KTR}}(\theta) = \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}m_{i,t}\cdot\min\left(r_{i,t}\cdot\hat{A}_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\cdot\hat{A}_{i,t}\right)\right], \tag{2}$$

where only tokens with $m_{i,t} = 1$, *i.e.*, those identified as important, contribute to the loss, allowing the model to focus supervision on semantically rich segments such as temporally anchored verbs, visually descriptive nouns, and discourse pivots with high uncertainty. As shown in Figure 2, the selected tokens differ across attribution strategies, reflecting their distinct roles in multimodal reasoning. In contrast, unselected tokens are largely low-information function words (*e.g.*, auxiliary verbs, pronouns, determiners, prepositions, and punctuation), as illustrated in Figure 6b. Their prevalence further validates the attribution mechanism's ability to filter redundant content and enhance learning efficiency.

Table 1: Performance comparison across reasoning-oriented and general-purpose video benchmarks, where * indicates scores evaluated by ourselves.

Models	Size	# Frames	Video Reasoning Benchmark			Video General Benchmark	
			Video-Holmes	VideoMMMU	MMVU(mc)	TempCompass	VideoMME
Proprietary MLLMs							
GPT-4o	–	–	42.0	61.2	75.4	73.8	71.9
GPT-5	–	–	46.7*	84.6	82.6*	83.3*	86.7
Gemini-1.5-Pro	–	–	41.3	53.4	71.2	67.1	75.0
Gemini-2.5-Pro	–	–	45.0	83.6	78.4*	84.3*	84.3
Open-Source MLLMs							
LLaVA-OV	7B	64	–	33.8	49.2	64.2	58.2
VILA-1.5	8B	64	–	33.8	49.2	58.8	58.2
Qwen2.5-VL	7B	–	27.8	47.4	59.2	67.9	**65.1**
Video-R1	7B	32	36.5	52.3	63.8	73.2	59.3
Video-RTS	7B	51.2	40.7	52.7	66.4	–	63.0
TW-GRPO	7B	16	32.9	51.3	65.8	73.3	55.1
SFT Models							
Qwen2.5-VL-SFT	7B	16	31.7	47.4	61.3	69.2	52.8
		32	33.9	49.4	63.5	69.9	55.4
		64	33.7	49.4	61.6	70.0	58.8
Video-KTR	7B	16	40.7	51.3	65.7	73.3	57.3
		32	41.6	52.6	65.9	73.4	60.3
		64	**42.7**	**53.1**	**66.6**	**73.5**	62.5

4 Experiments

4.1 Experimental Setup

Datasets. For RL data, we select the raw data released by Video-R1 (Feng et al., 2025), containing 260K samples with verifiable answers. We conduct strict difficulty sampling to enhance the overall difficulty and coverage of the training samples. Specifically, we remove samples that achieve above 80% or below 20% accuracy over 8 tails using Video-R1-SFT (Feng et al., 2025), leading to $\sim$ 15K high-quality subset. We also incorporate 1.5K training examples of Video-Holmes (Cheng et al., 2025) to form the final training set.

Implementations. The Video-KTR directly perform RL training on the Video-R1 SFT checkpoint, which itself is built upon the Qwen-2.5-VL-7B (Bai et al., 2025) backbone. To improve the efficiency, we limit the input to a maximum of 16 frames with a maximum resolution of $128 \times 28 \times 28$ pixels. We set the learning rate to $2e - 6$, the global batch size to 32, the rollout batch size to 256, the maximum sequence length to 16384; the KL divergence coefficient (β) is configured at 0.4; and 8 responses are sampled for each prompt with a temperature of 1.0. During the inference, we may extend the maximum frames to 64 with a maximum resolution of $256 \times 28 \times 28$.

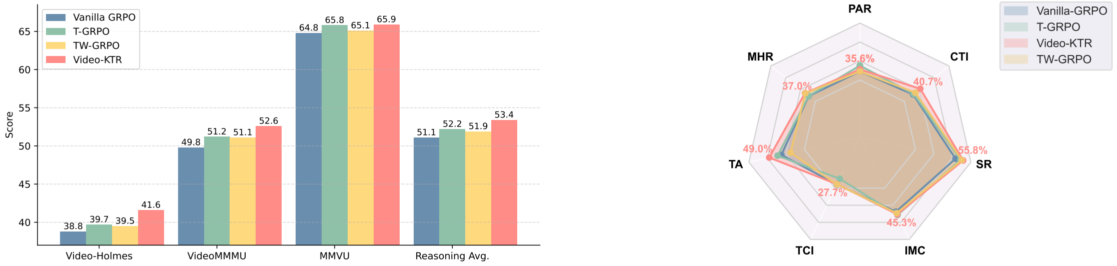
Benchmarks. To comprehensively assess the performance of Video-KTR, we select five challenging video benchmarks: (1) reasoning-oriented benchmarks, including **Video-Holmes** (Cheng et al., 2025), **VideoM-**

MMU (Hu et al., 2025), and **MMVU** (Zhao et al., 2025), to assess the temporal, causal, and knowledge-intensive reasoning capability of MLLMs; (2) general understanding benchmarks, including **TempCompass** (Liu et al., 2024) and **VideoMME** (Fu et al., 2025). Note Video-Holmes is a newly established benchmark that poses significant challenges, requiring models to integrate dispersed narrative cues from suspenseful short films and perform complex deductive reasoning beyond basic visual recognition.

Baselines. We compare Video-KTR against a diverse set of baselines, encompassing both general-purpose and reasoning-oriented MLLMs of similar size: LLaVA-OneVision (LLaVA-OV; Li et al. (2025a)), VILA-1.5 (Liu et al., 2025), Qwen2.5-VL (Bai et al., 2025), Video-R1 (Feng et al., 2025), Video-RTS (Wang et al., 2025d), and TW-GRPO (Dang et al., 2025).

4.2 Results and Analysis

Comparison with state-of-the-art models. We evaluate Video-KTR against a diverse set of video reasoning models on three reasoning and two general video understanding benchmarks. Reported in Table 1, Video-KTR consistently achieves state-of-the-art results. On **Video-Holmes**—which targets high-level temporal and social reasoning in short films—our model attains **42.7%**, surpassing all open-source baselines and closely matching proprietary models like GPT-5 (OpenAI (2025); 46.7%) and Gemini-2.5-Pro (Comanici et al. (2025); 45.0%). On knowledge-intensive benchmarks, *i.e.*, **MMVU(mc)** and **VideoMMMU**, Video-KTR achieves **66.6%** and **53.1%**, respectively, highlighting its strength in complex reasoning. Video-KTR also delivers competitive performance on general video understanding, showing that improved reasoning does not compromise broad comprehension. Notably, accuracy increases steadily with more input frames (16 to 64), underscoring the scalability and robustness of our token-aware RL for longer temporal sequences.



(a) Cross-Benchmark Comparison.

(b) Performance across Video-Holmes subtasks.

Figure 3: Performance comparison on (a) video reasoning benchmarks and (b) detailed subtasks of the Video-Holmes benchmark. Subtask abbreviations: **SR** = *Social Reasoning*, **IMC** = *Intention & Motive Chaining*, **TCI** = *Temporal Causal Inference*, **TA** = *Timeline Analysis*, **MHR** = *Multimodal Hint Reasoning*, **PAR** = *Physical Anomaly Reasoning*, **CTI** = *Core Theme Inference*.

Research Question (RQ1): How effective is Video-KTR compared with other video RL methods? To isolate the effect of training data and recipe, we conduct a controlled study where all models are trained under an identical dataset and training recipe. Depicted in Figure 3a, Video-KTR consistently outperforms vanilla GRPO, T-GRPO, and TW-GRPO, demonstrating its effectiveness independent of data quality or quantity. Figure 3b reports Video-Holmes results by subtask. Video-KTR achieves marked gains in Timeline Analysis (**TA**) and Core Theme Inference (**CTI**), tasks that require advanced temporal reasoning and holistic content understanding. The results confirm that our method enhances both temporal and structural reasoning, validating the token-aware training design. A qualitative comparison with T-GRPO (Appendix B.6) illustrates

Table 2: Performance of token selection using entropy (E), visual (V), and temporal (T) attributions.

Strategy	E	V	T	Video-Holmes	Video-MMMU	MMVU	Avg
Vanilla GRPO	✗	✗	✗	38.8	49.8	64.8	51.1
E	✓	✗	✗	39.5	49.2	64.3	51.0
V	✗	✓	✗	40.5	51.9	65.1	52.5
T	✗	✗	✓	**42.1**	50.1	65.5	52.6
V & T	✗	✓	✓	41.3	52.1	65.2	52.9
E & T	✓	✗	✓	39.6	50.9	64.3	51.6
V & E	✓	✓	✗	41.0	51.1	66.4	52.8
V & E & T	✓	✓	✓	41.6	**52.6**	**65.9**	**53.4**

that Video-KTR identifies key visual and temporal cues more effectively, leading to correct predictions and highlighting the benefit of token-level optimization in aligning model attention with critical signals.

RQ2: What is the impact of different attribution signals on video reasoning performance? To assess the impact of different attribution signals, we ablate the three token types—entropy-aware (E), visual-aware (V), and temporal-aware (T)—as shown in Table 2. Each signal individually outperforms vanilla GRPO, with temporal-aware tokens giving the largest gain on Video-Holmes, reflecting their importance for sequence reasoning. However, using temporal-aware tokens alone hurts performance on other benchmarks (e.g., Video-MMMU), suggesting that a single signal can bias the policy toward a narrow reasoning mode. Pairwise combinations yield further improvements, and the full E+V+T setup consistently achieves the best results across benchmarks. These findings show that the three signals are complementary and that jointly leveraging them enables more accurate and robust policy optimization.

RQ3: What are the linguistic divergences in token selection across attribution strategies? To assess the complementarity of our attribution strategies, we analyze the linguistic distribution and overlap of selected tokens. As shown in Figure 4, visual-aware tokens are mostly `NOUN`s (24.8%), consistent with their role in object grounding. Temporal-aware tokens emphasize `VERB`s (21.2%) and `PRON`s (11.0%), reflecting action transitions and temporal references. Entropy-aware tokens contain more `ADV`s (8.8%), capturing discourse modulation and uncertainty. These patterns suggest that modality-aware token selection leverages complementary reasoning dimensions—perceptual grounding, temporal structure, and uncertainty cues—to enhance video understanding.

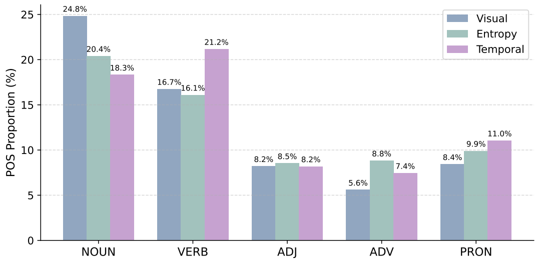


Figure 4: POS ratio across selection strategies.

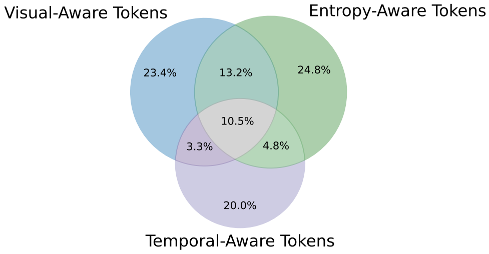
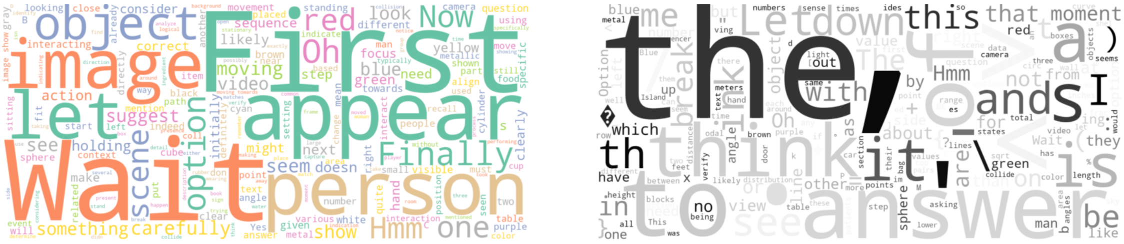


Figure 5: Distribution of updated tokens.

As illustrated in Figure 6a, the word cloud of selected tokens highlights semantically rich cues that directly contribute to multimodal reasoning, such as content words related to objects (*object*, *person*), actions (*appear*, *hold*), and reasoning markers (*wait*, *let*, *consider*, *finally*). These tokens capture critical visual and



(a) Word cloud of selected critical tokens.

(b) Word cloud of unselected low-value tokens.

Figure 6: Comparison of tokens selected as reasoning-critical and unselected tokens. Selected tokens concentrate on semantically informative content, whereas unselected tokens are dominated by function words, indicating that the selection mechanism emphasizes reasoning-critical tokens while filtering low-value ones.

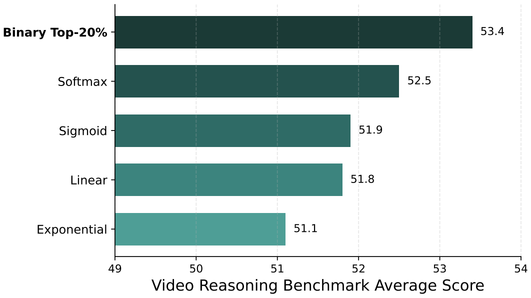


Figure 7: Benchmark Scores under Different Token Weighting Methods.

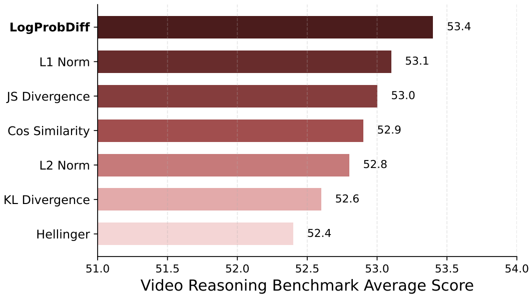


Figure 8: Benchmark Scores under Different Distance Calculation Methods.

temporal evidence, guiding the model to focus on informative regions of the input. In contrast, as shown in Figure 6b, unselected tokens are dominated by low-information function words, including auxiliary verbs (`AUX`: *is*, *have*), pronouns (`PRON`: *I*, *we*), determiners (`DET`: *the*, *this*), prepositions (`ADP`: *in*, *on*), and various punctuation or formatting symbols (*e.g.*, . , ,). Their prevalence confirms the effectiveness of our attribution strategies in filtering out syntactic noise while prioritizing semantically salient reasoning cues.

RQ4: How do token weighting strategies affect selective RL? We compare several token-weighting strategies to examine whether soft importance modulation can improve upon the hard selection used in Video-KTR. The methods include: (1) **Binary Top-20%**, which updates only the top 20% key tokens and masks the rest; (2) **Softmax Weighting**, which normalizes importance scores into a probability distribution; (3) **Sigmoid Weighting**, which applies a smooth nonlinear mapping; (4) **Linear Weighting**, which scales updates proportionally to token scores; and (5) **Exponential Weighting**, which accentuates score gaps. As shown in Figure 7, the hard top-20% strategy consistently outperforms all soft weighting variants, indicating that strict selection yields clearer credit assignment while soft modulation spreads updates over less relevant tokens. Token-level overlap analysis (Figure 5) further shows that, after applying the union of visual, temporal, and entropy signals, the final update ratio reaches about 40%, with most tokens unique to individual strategies. This demonstrates that each attribution signal captures a distinct and complementary subset of reasoning-relevant tokens.

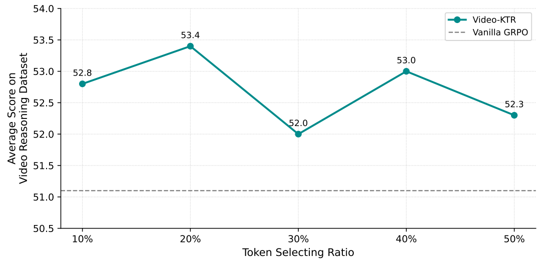


Figure 9: Effect of update ratios.

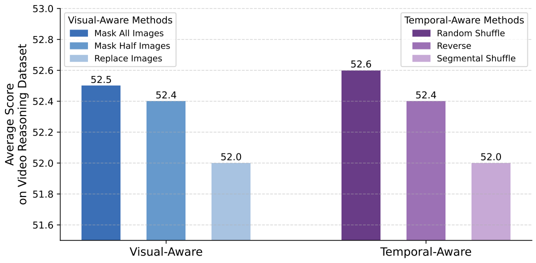


Figure 10: Effect of perturbation strategies.

RQ5: Do distributional shifts improve token attribution beyond log-probability? To identify informative tokens for reinforcement updates, we assess both changes in predicted log-probabilities and shifts in the full output-logit distribution under modality-specific perturbations. Specifically, we quantify token-wise distances between pre- and post-perturbation logits using metrics like L1/L2 norms, KL/JS divergences, cosine similarity, and Hellinger distance. Depicted in Figure 8, the simple log-probability difference consistently outperforms these alternatives. It is robust, easy to implement, and efficient-requiring only a subtraction of two log-probabilities per token-making it highly scalable. These results suggest log-probability differences as a reliable and efficient signal for tracking prediction-confidence shifts and identifying informative tokens.

RQ6: How does the token-level update ratio affect performance? We ablate the token-level update ratio—the proportion of key tokens receiving full gradient updates—with all token types applied jointly. As shown in Figure 9, accuracy follows a double-peak pattern and achieves the best performance at a 20% ratio. Performance drops at higher ratios (30–50%), indicating that moderate selective updating is optimal, while overly large update sets introduce noise and reduce stability.

RQ7: Do perturbation strategies influence token identification and model performance? To assess the robustness of our token identification method, we examine alternative perturbation strategies for visual and temporal dependencies. For visual perturbation, we consider *masking all frames* (default), *masking half*, and *replacing frames with unrelated content*. For temporal analysis, we compare *random shuffling* (default), *sequence reversal*, and *segmental shuffling*. In all cases, the token update ratio is fixed at 20%. Depicted in Figure 10, all variants achieve comparable performance, with the default settings—full-frame masking (52.5) and random shuffling (52.6)—performing slightly better. These results suggest that perturbation strength has limited impact, confirming the robustness of our method. We therefore adopt the default strategies for their simplicity and reliability, without requiring manual tuning.

5 Conclusion

We introduce Video-KTR, a modality-aware policy shaping framework that enhances video reasoning by shifting RL from coarse sequence supervision to selective, token-level attribution signal assignment. By integrating visual, temporal, and uncertainty-based attribution, Video-KTR focuses on semantically critical tokens, improving both accuracy and interpretability. Experiments across five benchmarks—highlighted by 42.7% on Video-Holmes, surpassing GPT-4o—demonstrate SOTA or highly competitive performance. Ablations confirm the complementary roles of three signals and robustness of targeted updates. The approach integrates seamlessly into RL and enhances reasoning without compromising general video understanding.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, and Peng Wang et al. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Deli Zhao, Anh Tuan Luu, and Yu Rong. GeoPQA: Bridging the visual perception gap in MLLMs for geometric reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25680–25688, Suzhou, China, November 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.1400/.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, XIAOJUAN QI, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling RL to long videos. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=TxedB8hI5O.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *ArXiv*, abs/2505.21374, 2025. URL https://arxiv.org/abs/2505.21374.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, and Nan-Jiang Jiang. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261, 2025. URL https://arxiv.org/abs/2507.06261.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *ArXiv*, abs/2505.24718, 2025. URL https://arxiv.org/abs/2505.24718.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, abs/2501.12948, 2025. URL https://arxiv.org/abs/2501.12948.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *ArXiv*, abs/2503.17352, 2025. URL https://arxiv.org/abs/2503.17352.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=a2JTVVvcEl.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.pdf.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *ArXiv*, abs/2501.13826, 2025. URL https://arxiv.org/abs/2501.13826.

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

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. URL https://arxiv.org/abs/1603.07396.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *ArXiv*, abs/2509.21268, 2025. URL https://arxiv.org/abs/2509.21268.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *ArXiv*, abs/2504.06958, 2025b. URL https://arxiv.org/abs/2504.06958.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.517/.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4122–4134, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Liu_NVILA_Efficient_Frontier_Visual_Language_Models_CVPR_2025_paper.pdf.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.177/.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *ArXiv*, abs/2503.07365, 2025. URL https://arxiv.org/abs/2503.07365.

OpenAI. Openai o1 system card. *ArXiv*, abs/2412.16720, 2024. URL https://arxiv.org/abs/2412.16720.

OpenAI. Introducing GPT-5. https://openai.com/blog/introducing-gpt-5, 2025.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive GRPO. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2YPypUwWIv.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *ArXiv*, abs/2504.07615, 2025. URL https://arxiv.org/abs/2504.07615.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun MA, and Chao Zhang. video-SALMONN-o1: Reasoning-enhanced audio-visual large language model. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=y62fhuA69I.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NdScoAix25.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *ArXiv*, abs/2506.13654, 2025. URL https://arxiv.org/abs/2506.13654.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025a. URL https://arxiv.org/abs/2504.08837.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=yfcpdY4gMP.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *ArXiv*, abs/2503.12605, 2025c. URL https://arxiv.org/abs/2503.12605.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-RTS: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28126–28140, Suzhou, China, November 2025d. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.1428/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chang Gao et al. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, Hao Zhang, and Yu Rong. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *ArXiv*, abs/2507.22607, 2025. URL https://arxiv.org/abs/2507.22607.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.pdf.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Zhang_R1-VL_Learning_to_Reason_with_Multimodal_Large_Language_Models_via_ICCV_2025_paper.pdf.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Zhao_MMVU_Measuring_Expert-Level_Multi-Discipline_Video_Understanding_CVPR_2025_paper.pdf.

A LIMITATIONS

While Video-KTR has demonstrated notable improvements and potential, several avenues remain open for further exploration. First, under challenging conditions such as low light, occlusion, rapid motion, or OCR/ASR noise, visual–text alignment may be affected. Enhancing the stability and robustness of key token selection in diverse scenarios represents a promising direction for future work. Second, Our experiments have primarily focused on video reasoning benchmarks, whereas broader applications such as video captioning or temporal grounding have not yet been systematically examined. Extending validation to these domains would offer valuable insights into the generalization of the framework. Third, the current framework emphasizes vision–language modeling without explicitly incorporating additional modalities such as audio or motion sensor data. Since these signals often provide complementary information, integrating richer multimodal inputs holds considerable potential for future research.

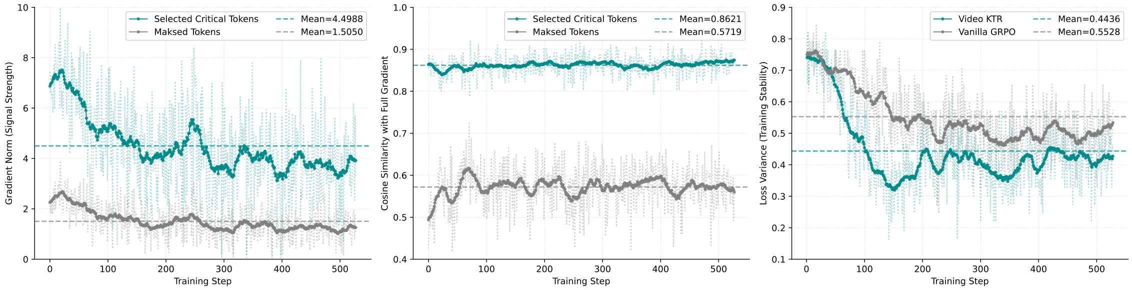


Figure 11: Video-KTR exhibits lower loss variance, stronger gradient signals, and higher gradient alignment during training. Due to the high variance of the accuracy curves, we smooth the curves using window smoothing with a window size of 20.

B More Results and Analysis

B.1 Gradient Decomposition and Noise Interpretation

To analyze whether masked tokens meaningfully affect optimization, we examine gradients at the final layer, where updates most directly influence policy learning. Let the sequence-level loss be $L = \sum_t \ell_t$ and denote the final-layer gradient for token t by

$$g_t = \nabla_{\theta^{(L)}} \ell_t = \frac{\partial \ell_t}{\partial h_t^{(L)}} \cdot \frac{\partial h_t^{(L)}}{\partial \theta^{(L)}},$$

where $h_t^{(L)}$ is the last-layer hidden state and $\theta^{(L)}$ denotes all parameters of the final layer. We partition tokens into a critical set $\mathcal{S}$ and a residual set (unselected low-value tokens) $\mathcal{R}$, yielding the following gradients

$$g_{\text{full}} = \sum_t g_t, \qquad g_{\text{KTR}} = \sum_{t \in \mathcal{S}} g_t, \qquad g_{\text{rest}} = \sum_{t \in \mathcal{R}} g_t.$$

To assess how these components contribute to parameter updates, we compare the directional agreement between each aggregated gradient and the full gradient. Rather than considering only their magnitudes, we evaluate whether two gradients point in similar directions in parameter space. This is measured using cosine similarity, written as $\cos(g_a, g_b)$, which reflects how much one gradient aligns with the other: values close to 1 indicate that g_a and g_b point in nearly the same direction and thus provide mutually reinforcing updates. In practice, we compute $\cos(g_{\text{KTR}}, g_{\text{full}})$ and $\cos(g_{\text{rest}}, g_{\text{full}})$ to quantify how much each component contributes to the effective update direction.

As shown in Figure 11, Video-KTR produces larger, cleaner, and more directionally consistent gradient updates compared to vanilla GRPO. The mean gradient norm of the selected critical tokens is 4.50, which is roughly three times higher than that of the masked tokens. Likewise, the directional agreement with the full gradient is substantially stronger: the cosine similarity of critical-token gradients averages 0.862, compared with 0.572 for the masked-token gradients. These results indicate that the selected tokens dominate the effective update magnitude and align far more closely with the model's underlying optimization trajectory. In contrast, gradients from residual masked tokens are both much weaker and directionally inconsistent, suggesting that they contribute little meaningful learning signal and instead introduce noisy perturbations into the optimization process.

Table 3: Performance comparison on image benchmarks. Results show that restricting updates to visual tokens already yields clear gains over GRPO variants, highlighting the importance of visual token optimization. Nevertheless, the full Video-KTR model achieves the best performance across all benchmarks, underscoring the additional benefits of integrating temporal and entropy-aware signals.

	MMMU(val)	**AI2D**	**MMSTAR**	**ChartQA**
Qwen2.5-VL-SFT	42.89	75.68	53.07	69.92
Vanilla GRPO	44.33	76.30	53.60	73.08
T-GRPO	44.67	77.51	54.32	77.04
KTR-Visual Tokens Only	46.44	79.27	57.38	77.08
Video-KTR	47.33	80.34	57.77	80.68

We further compare the training loss variance, which reflects overall optimization stability. Video-KTR achieves more stable training dynamics: its mean loss variance is 0.4436, lower than that of vanilla GRPO (0.5528) across training steps, supporting that masking low-information tokens reduces gradient noise and leads to more stable reinforcement learning updates.

Together, these results validate the effectiveness of our token-selection mechanism and highlight its role in guiding the model toward more meaningful and robust multimodal reasoning behavior.

B.2 Additional Evaluation on Image Benchmarks

To further assess the contribution of visual tokens, we introduce a variant of our method, denoted as KTR-Visual Tokens-Only, where reinforcement learning updates are restricted to visual-related tokens while discarding temporal and high-entropy token signals. This setting isolates the effect of visual supervision and provides a controlled comparison against the full Video-KTR framework. As shown in Table 3, this variant yields notable gains over baseline GRPO methods on several classic image understanding benchmarks, including MMMU-val (Yue et al., 2024), AI2D (Kembhavi et al., 2016), MMStar (Chen et al., 2024), and ChartQA (Masry et al., 2022), underscoring the importance of visual token optimization. Nevertheless, the complete Video-KTR achieves the best overall performance, suggesting that the integration of temporal and uncertainty-aware token signals further enhances cross-modal reasoning.

B.3 Computational Cost Analysis

Table 4: Comparison of training cost and forward efficiency. TT = training time, Mem = peak memory, SPS = samples per second, Lat. = latency. Video-KTR introduces only marginal overhead compared with Vanilla GRPO and T-GRPO.

	TT (h)	**Mem (GB)**	**SPS**	**fwd Lat. (s)**	**fwd FLOPs per GPU(T)**
Vanilla GRPO	4.7	77.9	0.946	4.70	116.6
T-GRPO	5.1	78.3	0.872	4.89	128.6
Video-KTR	5.2	78.5	0.855	4.92	129.9

Our experiments were conducted on 32 H100 (80GB) GPUs, and the training of Video-KTR was completed in 5.2 hours. As summarized in Table 4, the additional computational overhead introduced by our method remains modest. The key reason is that token importance scores are computed in a single extra forward pass, avoiding repeated per-token calculations. Compared with vanilla GRPO and T-GRPO, Video-KTR

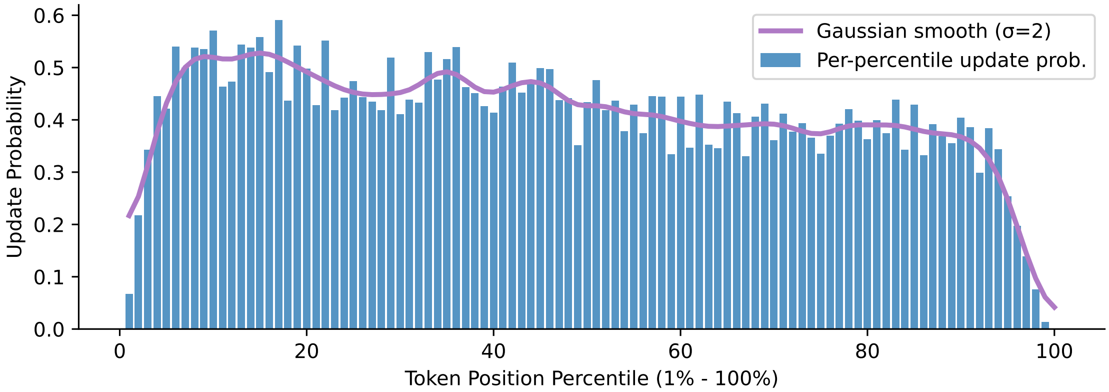


Figure 12: Distribution of updated token positions in Video-KTR. Early tokens show higher update probabilities, while later tokens still receive non-negligible updates, indicating that the model continues refining its reasoning rather than relying solely on pretrained linguistic priors.

shows almost no increase in peak memory usage (77.9 GB → 78.5 GB) and only a minor rise in forward latency (4.70 s → 4.92 s). The overall training time grows slightly from 4.7–5.1 hours to 5.2 hours, while forward FLOPs per GPU increase moderately (116.6 T → 129.9 T). These results indicate that our token-level reinforcement strategy delivers significant performance improvements at a marginal computational cost.

B.4 Updated Token Position Analysis

To better understand where reinforcement learning updates are applied, we analyze the distribution of updated token positions. As shown in Figure 12, tokens at earlier positions exhibit consistently higher update probabilities than those at later positions. This pattern is expected: early tokens play a larger role in determining the overall direction of the answer, whereas later outputs tend to be more deterministic and less influenced by input variability. Nevertheless, we also observe non-negligible updates in later positions, suggesting that the model does not merely rely on pretrained linguistic priors but continues to refine its reasoning based on the input content.

B.5 Generalize Video-KTR to Smaller MLLM

To assess generalizability beyond the 7B-scale backbone, we apply our training strategy to the smaller Qwen2.5-VL-3B model. As shown in Table 5, our method consistently outperforms vanilla GRPO across all frame settings (16, 32, and 64) and benchmarks. These findings demonstrate that our token-aware reinforcement learning strategy remains effective for smaller-scale MLLMs.

B.6 Qualitative Performance of Video-KTR

To qualitatively demonstrate Video-KTR's advantages, several representative examples from Video-Holmes are presented. In Figure 13, our model correctly identifies her ability to "ignore spatial physical limitations"—the official answer—by attending to essential visual and temporal cues, such as her "disappearance and reappearance" and "unpredictable spatial movements." Conversely, T-GRPO inaccurately attributes the rule to "transformation and disguise," underscoring its inability to ground reasoning effectively in the video's

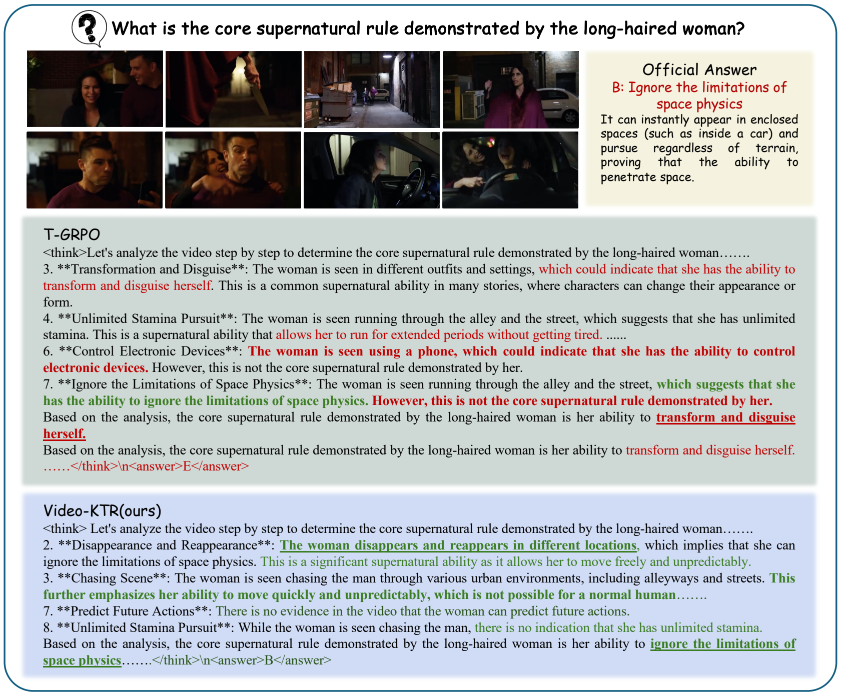
Table 5: Validation on Qwen2.5-VL-3B with different frame settings (16/32/64). Video-KTR consistently outperforms Vanilla GRPO, demonstrating robust improvements across benchmarks.

Models	**Frames**	**Video-Holmes**	**VideoMMMU**	**MMVU**	**TempCompass**	**VideoMME**
Vanilla GRPO	16	32	44.0	58.4	65.2	51.8
Video-KTR	16	33.2	45.1	60.0	67.2	53.0
Vanilla GRPO	32	33.1	44.7	59.5	66.7	54.1
Video-KTR	32	36.2	45.3	61.0	67.9	56.2
Vanilla GRPO	64	34.1	45.0	60.3	66.8	56.0
Video-KTR	64	36.3	46.8	61.3	68.0	57.5

actual content. This example illustrates the effectiveness of our token-level optimization in aligning model attention with critical visual and temporal signals. In Figure 14, Video-KTR further demonstrates its robustness by correctly inferring the underlying character relationship—that the two individuals are friends rather than brothers—whereas T-GRPO, lacking any supporting evidence, incorrectly selects a fraternal relationship. In Figure 15, Video-KTR accurately reasons about the character's motivation, identifying that the man pretends to be dead to escape the woman's attempted murder. In contrast, T-GRPO misidentifies the yellow gloves as a banana, leading it to the wrong conclusion that the man was awakened by the woman's phone call. In Figure 16, Video-KTR successfully tracks the temporal progression in a chaotic chase scene, recognizing that the man first tries to retreat into the room, fails, and is then relentlessly pursued. T-GRPO, however, mistakenly assumes that the man is chased continuously before attempting to return to the room, thereby misrepresenting the actual event order. These examples demonstrate Video-KTR's enhanced capacity for temporal modeling and consistent logical reasoning, allowing it to capture complex event structures with greater reliability.

C LLM Usage Statement

In accordance with the ICLR 2026 policy on large language model (LLM) usage, we disclose that an LLM (OpenAI's ChatGPT) was used in a **limited and non-substantive capacity**. Specifically, it assisted with *minor language editing* (*e.g.*, grammar and phrasing), provided suggestions on figure color choices, and offered LaTeX formatting guidance. **No research ideas, experimental designs, analyses, or substantive scientific writing were generated by the LLM**. All scientific contributions and intellectual content are entirely the work of the authors.



What is the core supernatural rule demonstrated by the long-haired woman?

Official Answer

B: Ignore the limitations of space physics

It can instantly appear in enclosed spaces (such as inside a car) and pursue regardless of terrain, proving that the ability to penetrate space.

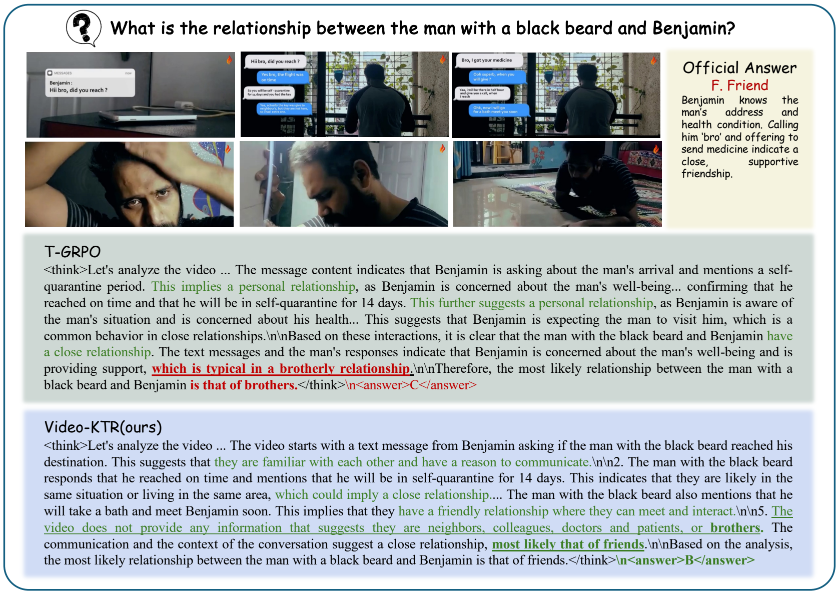
T-GRPO

<think>Let's analyze the video step by step to determine the core supernatural rule demonstrated by the long-haired woman…….
3. **Transformation and Disguise**: The woman is seen in different outfits and settings, which could indicate that she has the ability to transform and disguise herself. This is a common supernatural ability in many stories, where characters can change their appearance or form.
4. **Unlimited Stamina Pursuit**: The woman is seen running through the alley and the street, which suggests that she has unlimited stamina. This is a supernatural ability that allows her to run for extended periods without getting tired. ......
6. **Control Electronic Devices**: The woman is seen using a phone, which could indicate that she has the ability to control electronic devices. However, this is not the core supernatural rule demonstrated by her.
7. **Ignore the Limitations of Space Physics**: The woman is seen running through the alley and the street, which suggests that she has the ability to ignore the limitations of space physics. However, this is not the core supernatural rule demonstrated by her.
Based on the analysis, the core supernatural rule demonstrated by the long-haired woman is her ability to transform and disguise herself.
Based on the analysis, the core supernatural rule demonstrated by the long-haired woman is her ability to transform and disguise herself. ……</think>\n<answer>E</answer>

Video-KTR(ours)

<think> Let's analyze the video step by step to determine the core supernatural rule demonstrated by the long-haired woman…….
2. **Disappearance and Reappearance**: The woman disappears and reappears in different locations, which implies that she can ignore the limitations of space physics. This is a significant supernatural ability as it allows her to move freely and unpredictably.
3. **Chasing Scene**: The woman is seen chasing the man through various urban environments, including alleyways and streets. This further emphasizes her ability to move quickly and unpredictably, which is not possible for a normal human…….
7. **Predict Future Actions**: There is no evidence in the video that the woman can predict future actions.
8. **Unlimited Stamina Pursuit**: While the woman is seen chasing the man, there is no indication that she has unlimited stamina.
Based on the analysis, the core supernatural rule demonstrated by the long-haired woman is her ability to ignore the limitations of space physics…….</think>\n<answer>B</answer>

Figure 13: Qualitative comparison on Video-Holmes-Physical Anomaly Reasoning. Video-KTR infers the supernatural rule from temporal and visual cues, while T-GRPO produces an incorrect path.



What is the relationship between the man with a black beard and Benjamin?

Official Answer

F. Friend

Benjamin knows the man's address and health condition. Calling him 'bro' and offering to send medicine indicate a close, supportive friendship.

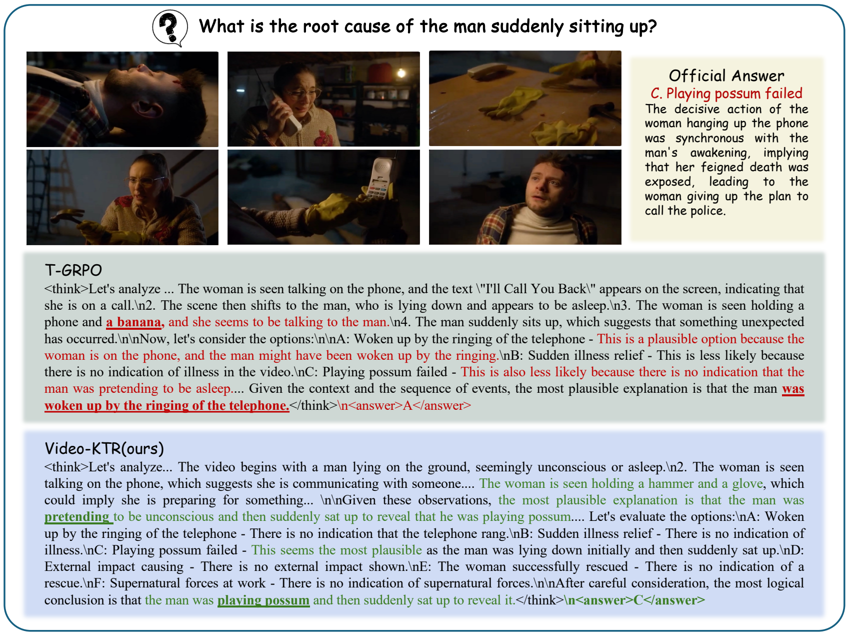
T-GRPO

<think>Let's analyze the video ... The message content indicates that Benjamin is asking about the man's arrival and mentions a self-quarantine period. This implies a personal relationship, as Benjamin is concerned about the man's well-being... confirming that he reached on time and that he will be in self-quarantine for 14 days. This further suggests a personal relationship, as Benjamin is aware of the man's situation and is concerned about his health... This suggests that Benjamin is expecting the man to visit him, which is a common behavior in close relationships.\n\nBased on these interactions, it is clear that the man with the black beard and Benjamin have a close relationship. The text messages and the man's responses indicate that Benjamin is concerned about the man's well-being and is providing support, which is typical in a brotherly relationship.\n\nTherefore, the most likely relationship between the man with a black beard and Benjamin is that of brothers.</think>\n<answer>C</answer>

Video-KTR(ours)

<think>Let's analyze the video ... The video starts with a text message from Benjamin asking if the man with the black beard reached his destination. This suggests that they are familiar with each other and have a reason to communicate.\n\n2. The man with the black beard responds that he reached on time and mentions that he will be in self-quarantine for 14 days. This indicates that they are likely in the same situation or living in the same area, which could imply a close relationship.... The man with the black beard also mentions that he will take a bath and meet Benjamin soon. This implies that they have a friendly relationship where they can meet and interact.\n\n5. The video does not provide any information that suggests they are neighbors, colleagues, doctors and patients, or **brothers.** The communication and the context of the conversation suggest a close relationship, **most likely that of friends**.\n\nBased on the analysis, the most likely relationship between the man with a black beard and Benjamin is that of friends.</think>**\n<answer>B</answer>**

Figure 14: Qualitative comparison on Video-Holmes-Social Reasoning.Video-KTR correctly inferred the underlying character relationship, whereas T-GRPO chose an answer without factual grounding.



What is the root cause of the man suddenly sitting up?

Official Answer
C. Playing possum failed
The decisive action of the woman hanging up the phone was synchronous with the man's awakening, implying that her feigned death was exposed, leading to the woman giving up the plan to call the police.

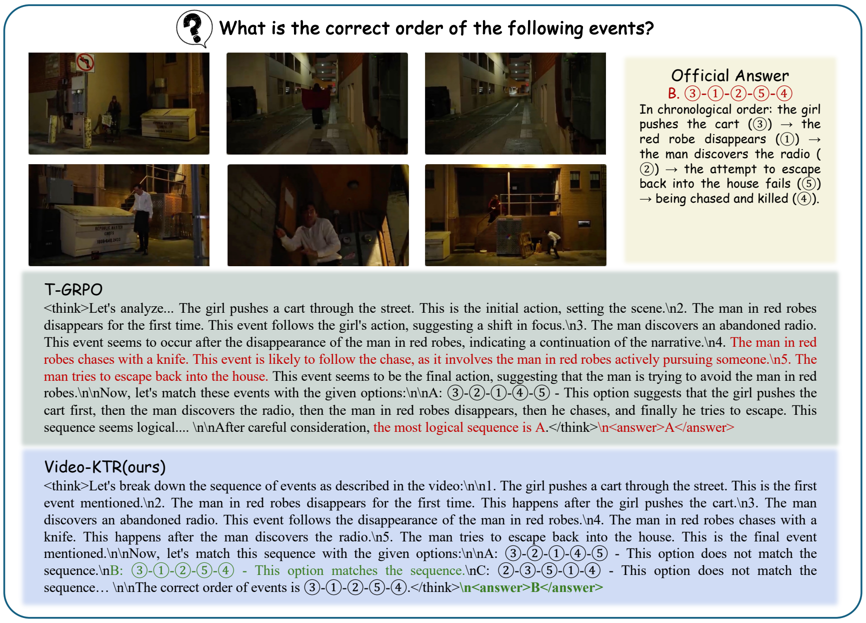
T-GRPO

<think>Let's analyze ... The woman is seen talking on the phone, and the text \"I'll Call You Back\" appears on the screen, indicating that she is on a call.\n2. The scene then shifts to the man, who is lying down and appears to be asleep.\n3. The woman is seen holding a phone and **a banana,** and she seems to be talking to the man.\n4. The man suddenly sits up, which suggests that something unexpected has occurred.\n\nNow, let's consider the options:\n\nA: Woken up by the ringing of the telephone - This is a plausible option because the woman is on the phone, and the man might have been woken up by the ringing.\nB: Sudden illness relief - This is less likely because there is no indication of illness in the video.\nC: Playing possum failed - This is also less likely because there is no indication that the man was pretending to be asleep.... Given the context and the sequence of events, the most plausible explanation is that the man **was woken up by the ringing of the telephone.**</think>\n<answer>A</answer>

Video-KTR(ours)

<think>Let's analyze... The video begins with a man lying on the ground, seemingly unconscious or asleep.\n2. The woman is seen talking on the phone, which suggests she is communicating with someone.... The woman is seen holding a hammer and a glove, which could imply she is preparing for something... \n\nGiven these observations, the most plausible explanation is that the man was **pretending** to be unconscious and then suddenly sat up to reveal that he was playing possum.... Let's evaluate the options:\nA: Woken up by the ringing of the telephone - There is no indication that the telephone rang.\nB: Sudden illness relief - There is no indication of illness.\nC: Playing possum failed - This seems the most plausible as the man was lying down initially and then suddenly sat up.\nD: External impact causing - There is no external impact shown.\nE: The woman successfully rescued - There is no indication of a rescue.\nF: Supernatural forces at work - There is no indication of supernatural forces.\n\nAfter careful consideration, the most logical conclusion is that the man was **playing possum** and then suddenly sat up to reveal it.</think>**\n<answer>C</answer>**

Figure 15: Qualitative comparison on Video-Holmes-Temporal Causal Inference. Video-KTR correctly reasoned about the character's motivation, whereas T-GRPO misidentified the yellow gloves as a banana and produced an incorrect answer.



What is the correct order of the following events?

Official Answer
B. ③-①-②-⑤-④
In chronological order: the girl pushes the cart (③) → the red robe disappears (①) → the man discovers the radio (②) → the attempt to escape back into the house fails (⑤) → being chased and killed (④).

T-GRPO

<think>Let's analyze... The girl pushes a cart through the street. This is the initial action, setting the scene.\n2. The man in red robes disappears for the first time. This event follows the girl's action, suggesting a shift in focus.\n3. The man discovers an abandoned radio. This event seems to occur after the disappearance of the man in red robes, indicating a continuation of the narrative.\n4. The man in red robes chases with a knife. This event is likely to follow the chase, as it involves the man in red robes actively pursuing someone.\n5. The man tries to escape back into the house. This event seems to be the final action, suggesting that the man is trying to avoid the man in red robes.\n\nNow, let's match these events with the given options:\n\nA: ③-②-①-④-⑤ - This option suggests that the girl pushes the cart first, then the man discovers the radio, then the man in red robes disappears, then he chases, and finally he tries to escape. This sequence seems logical.... \n\nAfter careful consideration, the most logical sequence is A.</think>\n<answer>A</answer>

Video-KTR(ours)

<think>Let's break down the sequence of events as described in the video:\n\n1. The girl pushes a cart through the street. This is the first event mentioned.\n2. The man in red robes disappears for the first time. This happens after the girl pushes the cart.\n3. The man discovers an abandoned radio. This event follows the disappearance of the man in red robes.\n4. The man in red robes chases with a knife. This happens after the man discovers the radio.\n5. The man tries to escape back into the house. This is the final event mentioned.\n\nNow, let's match this sequence with the given options:\n\nA: ③-②-①-④-⑤ - This option does not match the sequence.\nB: ③-①-②-⑤-④ - This option matches the sequence.\nC: ②-③-⑤-①-④ - This option does not match the sequence… \n\nThe correct order of events is ③-①-②-⑤-④.</think>**\n<answer>B</answer>**

Figure 16: Qualitative comparison on Video-Holmes-Timeline Analysis. Video-KTR correctly identified the sequence of events, whereas T-GRPO confused the temporal order during the chaotic chase.