

Video-KTR: Reinforcing Video Reasoning via Key Token Attribution

Ziyue Wang^{1,2}, Sheng Jin¹, Zhongrong Zuo¹, Jiawei Wu³, Han Qiu¹, Qi She¹, Hao Zhang^{4,*}, Xudong Jiang^{2,*}

¹ByteDance, ²School of Electrical and Electronic Engineering, Nanyang Technological University, ³National University of Singapore, ⁴College of Computing and Data Science, Nanyang Technological University
*Correspondence to Hao Zhang and Xudong Jiang.

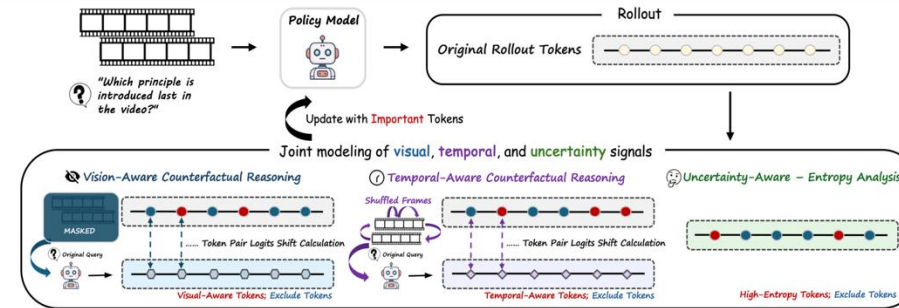


Introduction

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 the fine-grained links among visual inputs, temporal dynamics, and linguistic outputs, which limits 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: visual-aware tokens identified via counterfactual masking to reveal perceptual dependence, temporal-aware tokens detected through frame shuffling to expose temporal sensitivity, and 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—reaching 42.7% on Video-Holmes and surpassing GPT-4o—while delivering consistent gains on both reasoning and general video understanding tasks. Ablation studies further verify the complementary roles of the attribution signals and the robustness of targeted token-level updates. Overall, Video-KTR improves both accuracy and interpretability, offering a simple, drop-in extension to RL for complex video reasoning. Our code, data and models are all open-sourced on GitHub.

Methodology

Video-KTR identifies visual-, temporal-, and uncertainty-sensitive tokens, then updates only those key tokens during RL. Three complementary token sets are computed and their union defines the update mask.



Visual-aware Mask the video and measure logit shifts. Large changes identify tokens grounded in appearance or objects.
Temporal-aware Shuffle frame order and re-score tokens. Large changes identify tokens tied to event order or causality.
Entropy-aware Select high-entropy tokens that mark uncertainty, discourse pivots, or decision points.



Updated Important Tokens

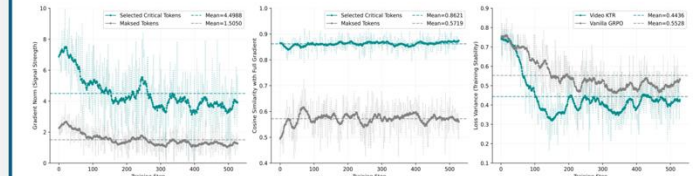
Dropped Tokens

RL Algorithms Ablation

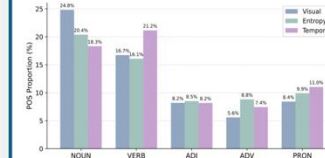
| Models | Size | # Frames | Video Reasoning Benchmark | | | | Video General Benchmark | |
|-------------------|------|----------|---------------------------|----------|----------|-------------|-------------------------|--|
| | | | Video-Holmes | VideoMMU | MMVU(mc) | TempCompass | VideoMME | |
| Proprietary 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 | |

Ablations & Insights

To analyze whether masked tokens meaningfully affect optimization, we examine gradients at the final layer, where updates most directly influence policy learning. 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 $\cos(g_{KTR}, g_{full})$ and $\cos(g_{rest}, g_{full})$ to quantify how much each component contributes to the effective update direction.

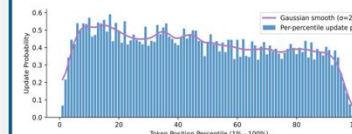


Video-KTR produces larger, cleaner, and more directionally consistent gradient updates compared to vanilla GRPO. The mean gradient norm of the selected critical tokens is 4.50, which is roughly three times higher than that of the masked tokens. Likewise, the directional agreement with the full gradient is substantially stronger: the cosine similarity of critical-token gradients averages 0.862, compared with 0.572 for the masked-token gradients.



| 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 | 50.1 | 65.5 | 52.6 |
| T | ✓ | ✓ | ✓ | 42.1 | 50.1 | 65.5 | 52.9 |
| 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 |

To assess the complementarity of our attribution strategies, we analyze the linguistic distribution and overlap of selected tokens. Visual-aware tokens are mostly NOUNs, consistent with their role in object grounding. Temporal-aware tokens emphasize VERBs and PRONs, reflecting action transitions and temporal references. Entropy-aware tokens contain more ADVs, capturing discourse modulation and uncertainty. These patterns suggest that modality-aware token selection leverages complementary reasoning dimensions—perceptual grounding, temporal structure, and uncertainty cue to enhance video understanding.



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.

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 and Gemini-2.5-Pro. On knowledge intensive benchmarks, i.e., MMVU and Video MMMU, 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

Our experiments were conducted on 32 H100 (80GB) GPUs.

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



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

The correct answer is B: Ignore the limitations of space physics

T-GRPO

<think>Let's analyze the video step by step to determine the core supernatural rule demonstrated by the long-haired woman.....
1. **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.
2. **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.....
3. **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.
4. **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><in>answer</in></answer>

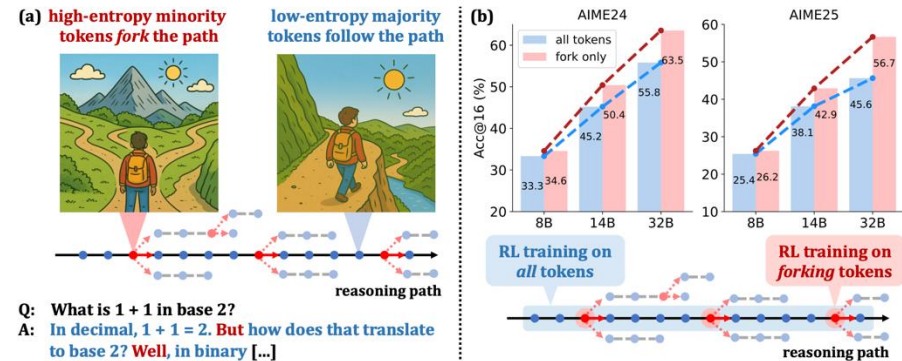
Video-KTR(ours)

<think>Let's analyze the video step by step to determine the core supernatural rule demonstrated by the long-haired woman.....
1. **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.....

Motivation & Challenge in Video Reasoning RL

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.

Many approaches overlook explicit temporal dependencies and causal structures—essential for effective reasoning. 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.



Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning

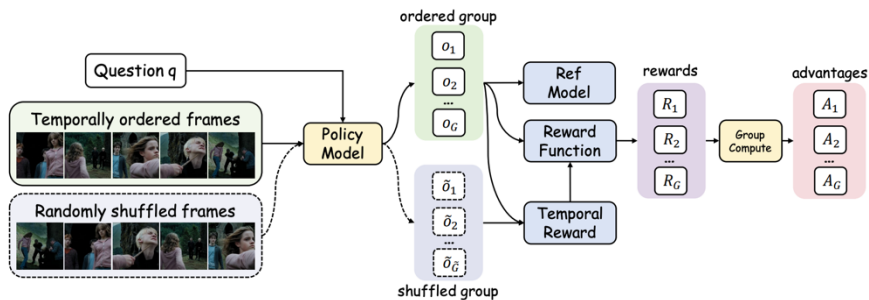
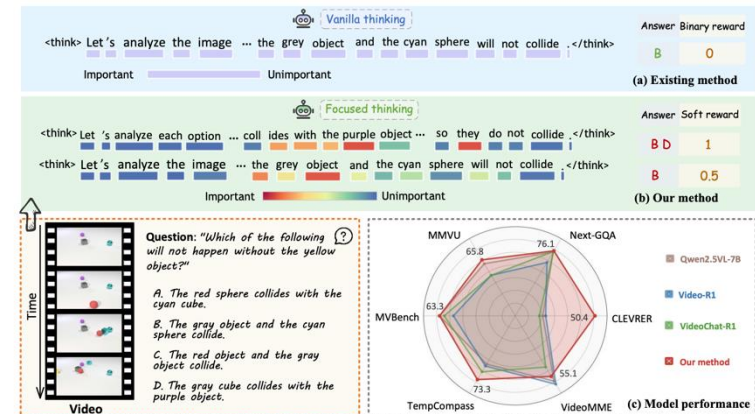


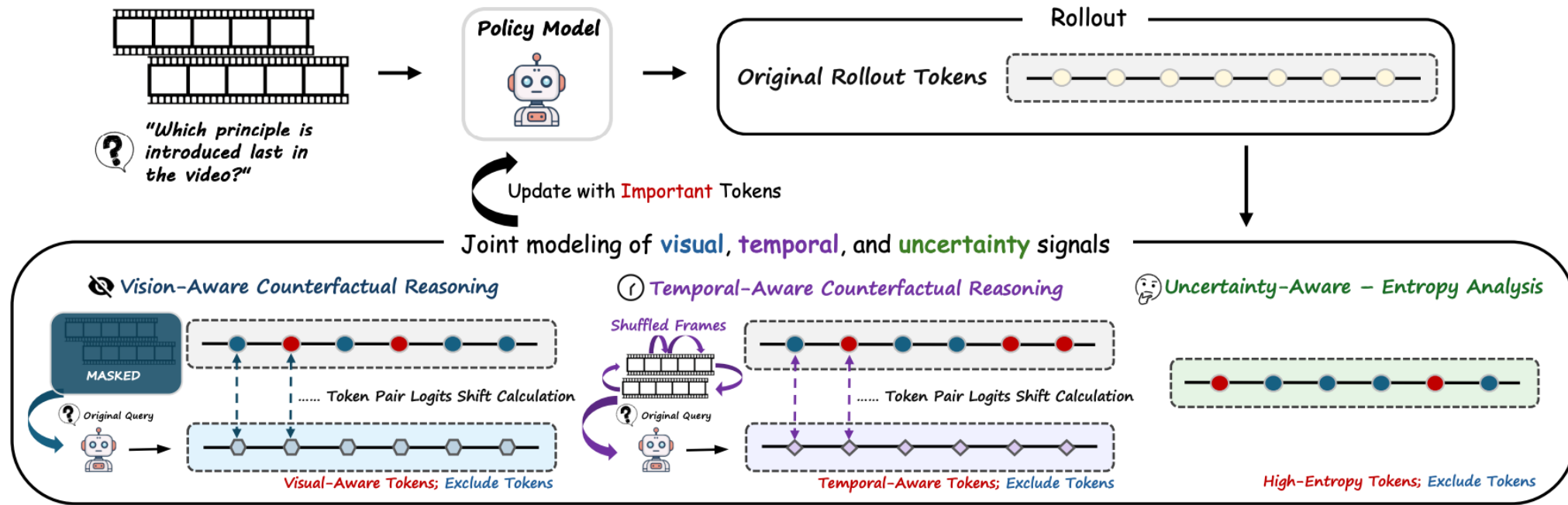
Figure 3: An illustration of our proposed T-GRPO algorithm.

Video-R1: Reinforcing Video Reasoning in MLLMs



Reinforcing Video Reasoning with Focused Thinking

Video-KTR Methodology



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: 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.

Temporal-aware Tokens: 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.

Uncertainty-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 as a logic-aware criterion to identify low-confidence tokens.

Video-KTR Performance

| 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 |

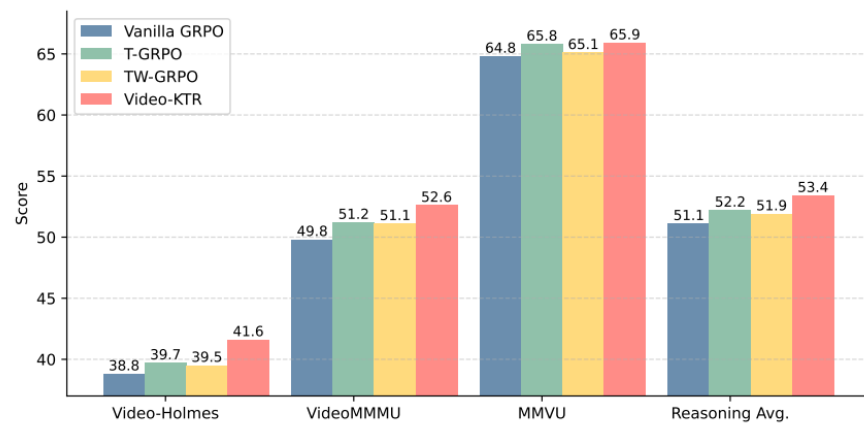
We evaluate Video-KTR against a diverse set of video reasoning models on three reasoning and two general video understanding benchmarks.

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 and Gemini-2.5-Pro.

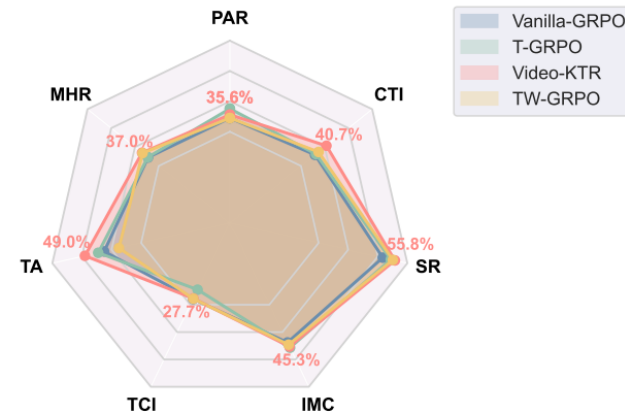
On knowledge-intensive benchmarks, i.e., MMVU(mc) and VideoMMMU, VideoKTR achieves 66.6% and 53.1%, respectively, highlighting its strength in complex reasoning.

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.

Ablations on RL Methods



(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*.

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.

Ablations on Different Attribution Signals

| 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 |

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. Pairwise combinations yield further improvements, and the full E+V+T setup consistently achieves the best results across benchmarks.

Analysis on Selective Updating Effectiveness

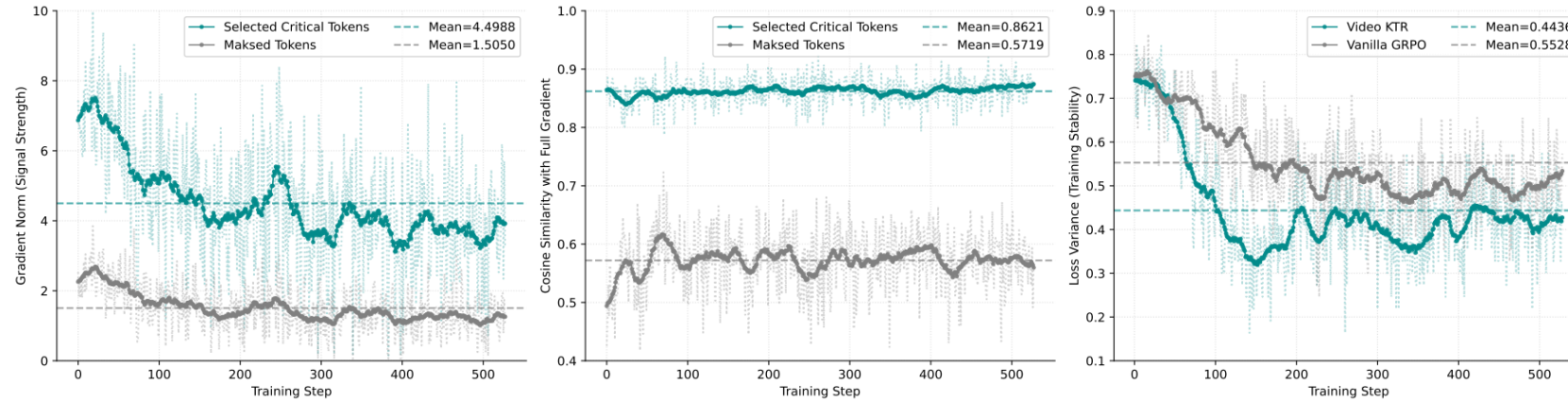


Figure 11: Video-KTR exhibits lower loss variance, stronger gradient signals, and higher gradient alignment during training. Due to the high variance of the accuracy curves, we smooth the curves using window smoothing with a window size of 20.

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.

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.

Thanks for your watching

[Ziyue Wang: ziyue005@e.ntu.edu.sg](mailto:ziyue005@e.ntu.edu.sg)

[Hao Zhang: hzhang26@outlook.com](mailto:h Zhang: hzhang26@outlook.com)

[Xudong Jiang: exdjiang@ntu.edu.sg](mailto:exdjiang@ntu.edu.sg)