LEARNING WHEN TO LOOK: ATTENTION-PROPOSED TEMPORAL REWARDS FOR GROUNDED VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video reasoning is a core capability of multimodal large language models (MLLMs), crucial for downstream tasks such as video question answering and captioning. Recent work has advanced this capability through reinforcement learning (RL), following its success in the language domain. However, video reasoning poses distinct challenges. A model must first locate the segments relevant to the question along the temporal axis, then reason from the visual evidence they provide. Existing rewards supervise only the final answer, leaving the first of these unconstrained. Recent approaches annotate temporal spans or invoke external perception tools at inference, incurring annotation or computational cost, and even an explicit interval reward leaves the fraction of correct answers unsupported by the model's own attention unchanged. We propose VideoAPR, a reinforcement learning framework that grounds video reasoning in the model's own attention. We introduce AP-GRPO, which derives a temporal reward from where the policy already attends, requiring no temporal annotation during training. Within each rollout, position-debiased attention proposes a candidate temporal segment, and the response the policy has already produced is rescored with the visual context restricted to that segment and to an equal-length control window elsewhere in the clip, with the margin between the two forming the reward. Verification is performed only during training, so inference remains a single forward pass at no additional cost. Extensive experiments demonstrate that VideoAPR achieves state-of-the-art performance across diverse video reasoning benchmarks, outperforming recent methods.

1 INTRODUCTION

Video understanding is a fundamental problem in computer vision, requiring models to perceive objects, actions, and their relations as they unfold over time. Multimodal large language models (MLLMs) (Bai et al., 2025; Li et al., 2024a; Hurst et al., 2024) have substantially strengthened the joint processing of visual and textual signals, opening new opportunities for complex video reasoning. Recently, reinforcement learning (RL) has emerged as an effective post-training paradigm for eliciting such reasoning. OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) show that optimizing against rule-based outcome rewards, without any process-level supervision, elicits long chain-of-thought reasoning in LLMs and generalizes better than supervised fine-tuning. The underlying algorithm, Group Relative Policy Optimization (GRPO) (Shao et al., 2024), estimates advantages from a group of sampled rollouts and thus requires no separate value model, making it readily applicable to multimodal settings. Motivated by this success, a growing body of work extends GRPO post-training to MLLMs, first on images (Liu et al., 2025; Shen et al., 2025) and then on videos (Feng et al., 2026; Wang et al., 2026a; Li et al., 2025).

However, models often arrive at the correct answer while attending to moments that do not support it, relying instead on scene-level appearance or on language priors over the question and its options (Yu et al., 2025). This mismatch between the answer and the evidence that actually produced it limits both the reliability and the interpretability of video reasoning. To address this, recent work makes the evidence explicit during reasoning. Some methods reward temporal order sensitivity (Feng et al., 2026) or semantic consistency between the generated rationale and the video (Wang et al., 2026a),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

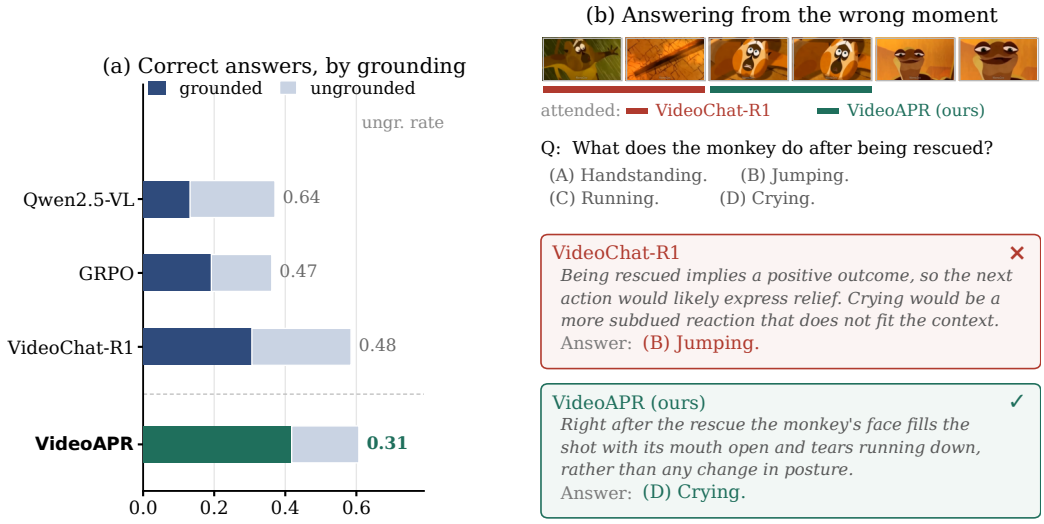

Figure 1: (a) Correct predictions split by whether the interval extracted from the model's attention agrees with the interval it predicts, with the ungrounded rate at the end of each bar. (b) A question the baseline answers from a moment that does not support it.

others require the policy to emit timestamps and score them with interval IoU against annotated spans (Li et al., 2025; Wang et al., 2026b), and more recent formulations equip the model with perception tools that resample frames on demand (Zhang et al., 2026; Lee et al., 2026). By exposing evidence, the reasoning process becomes traceable and verifiable.

Despite this progress, two challenges remain. The first is supervision. Interval IoU rewards require ground-truth temporal spans, which are costly to annotate and ambiguous at boundaries. The second lies in what these rewards constrain. Outcome rewards constrain only the answer (Xiao et al., 2026), and interval IoU constrains only what the model says about time, so neither requires the answer to follow from the moment the model actually attended to. Recent work shows that a sparse subset of attention heads encodes which segment a video MLLM is relying on (Du et al., 2026), giving us a way to measure this directly. As shown in Figure 1, comparing the window a model attends to against the window it predicts, we find that RL lowers the fraction of correct predictions its own attention does not support from 0.64 for the untrained backbone to about 0.47, and then stalls there. A method trained with an explicit interval IoU reward is no exception, raising grounded and ungrounded predictions in the same proportion rather than shifting the proportion itself. In short, grounding supervision should verify that the answer is recoverable from the evidence the model used, rather than that a predicted interval matches an annotation.

To address these challenges, we propose VideoAPR, a reinforcement learning framework that rewards a policy for answering from the evidence it attends to. Rather than supervising a predicted interval against an annotation, VideoAPR verifies the attended interval by intervention. If a segment genuinely supports an answer, that answer should remain recoverable once everything else is removed. For each question we extract the attended window from prefill attention, restrict the video to it, and score the response the policy produced under that restriction, comparing this against the score obtained on a window of identical size placed as far away as possible along the timeline. The margin between the two log-likelihoods forms the reward, granted only when the policy answered correctly on the full video, since crediting a response whose answer collapses under cropping would reward fragility rather than evidence. Matching the two window sizes makes the reward a comparison rather than an absolute score, so the policy cannot raise it by widening the interval, and any competence that transfers to arbitrary crops cancels between the two terms. No ground-truth spans are used at any point in training.

Extensive experiments across diverse video reasoning benchmarks (Yang et al., 2025; Hu et al., 2026; Zhao et al., 2025; Li et al., 2024b; Liu et al., 2024; Fu et al., 2025) demonstrate that VideoAPR outperforms existing methods and achieves state-of-the-art performance in video reasoning.

Our contributions are summarized as follows:

- We analyze whether RL-trained MLLMs answer from the moments they attend to, and show that existing rewards, including an explicit interval IoU reward, leave the fraction of ungrounded correct predictions unchanged.
- We propose VideoAPR, which verifies the attended interval by restricting the video to it and rewarding the margin over an equal-size distant window under correctness gating, requiring no ground-truth temporal spans.
- We demonstrate the effectiveness of VideoAPR across six video reasoning and understanding benchmarks, where it outperforms existing methods and achieves state-of-the-art performance.

2 Related Work

Reinforcement Learning for Video Reasoning. Following the success of outcome-reward RL in language models (Guo et al., 2025; Shao et al., 2024), a growing body of work applies Group Relative Policy Optimization to video MLLMs (Feng et al., 2026; Wang et al., 2026a; Li et al., 2025). Video-R1 (Feng et al., 2026) contrasts rollouts on ordered and shuffled frames to encourage temporal sensitivity, while VideoRFT (Wang et al., 2026a) adds a semantic-consistency term between the generated rationale and averaged frame features. A parallel line targets temporal alignment directly, rewarding interval IoU against annotated spans (Li et al., 2025; Wang et al., 2026b). More recent formulations move supervision into the inference procedure itself, equipping the policy with perception tools that resample frames on demand (Zhang et al., 2026; Lee et al., 2026). Despite their differences, these rewards evaluate the answer, the ordering of frames, or an interval the model states, and prior work has observed that a correct answer may still rest on irrelevant moments (Yu et al., 2025). In contrast, we reward a policy for answering from the evidence it attends to, verified by restricting the visual context rather than by matching a stated interval to an annotation.

Evidence Grounding in Video Reasoning. To make reasoning traceable, recent methods require models to expose the visual evidence they rely on. In image understanding, emitting boxes, coordinates, or localized crops has been shown to improve faithfulness by anchoring intermediate steps in observable regions (Sarch et al., 2026; Zheng et al., 2026; Shen et al., 2025). Video extensions incorporate timestamps, keyframes, and spatial grounding into the reasoning trace, and reward the exposed evidence directly (Meng et al., 2025; Wang et al., 2024; 2026b). Orthogonally, recent analysis shows that video MLLMs already encode temporal evidence internally, with a sparse subset of attention heads concentrating on the supporting segment during prefill even when the emitted timestamp is wrong, and exploits this at inference without updating the model (Du et al., 2026). We build on this line by turning the extraction into a training signal, so that grounding is optimized rather than corrected after the fact, and by verifying the attended interval through intervention.

3 Method

Figure 2 shows an overview of VideoAPR. After briefly reviewing GRPO (Section 3.1), we extract the interval a policy attends to from its own prefill attention and use it to measure whether RL-trained models answer from the moments they attend to (Section 3.2). We then turn the extraction into a reward by verifying that interval against an equal-length control window (Section 3.3).

3.1 Preliminaries

Group Relative Policy Optimization. Given a video V and a question q, GRPO (Shao et al., 2024) samples a group of G responses $\{o_1, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and scores each with a rule-based reward R_i. Rather than learning a separate value function, it estimates the advantage by

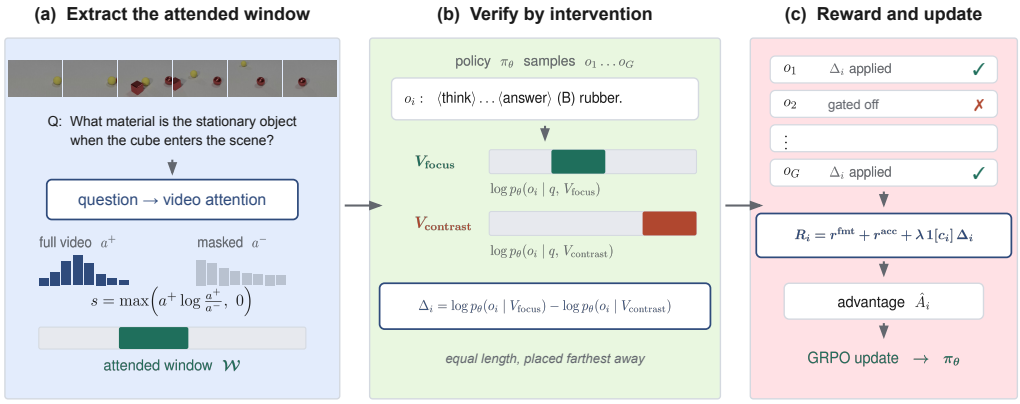

Figure 2: **Overview of VideoAPR.** (a) Before generation, a single prefill pass extracts question-to-video attention from a calibrated set of heads, contrasts it against a masked pass to remove position bias, and thresholds the resulting profile into an attended window $\mathcal{W}$. (b) Each rollout is rescored with the visual context restricted to $\mathcal{W}$ and to an equal-length control window placed farthest along the timeline, and the margin between the two log-likelihoods forms the verification signal. (c) The margin is added to the format and accuracy rewards under correctness gating, and the policy is updated with GRPO.

normalizing rewards within the group,

$$\hat{A}_i = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^{G}\right)}{\text{std}\left(\{R_j\}_{j=1}^{G}\right) + \epsilon}, \tag{1}$$

and optimizes the clipped objective

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i \hat{A}_i,\ \text{clip}\left(\rho_i, 1-\varepsilon, 1+\varepsilon\right) \hat{A}_i\right)\right] - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}), \tag{2}$$

where $\rho_i = \pi_\theta(o_i \mid q, V)/\pi_{\theta_{\text{old}}}(o_i \mid q, V)$ and π_{ref} denotes the frozen initial policy. In existing video RL, R_i combines a format reward with answer correctness and, for grounding-oriented tasks, an interval IoU term computed against an annotated span. Since advantages are group-relative, a reward term that takes the same value across all G rollouts cancels in Equation 1 and contributes no gradient.

3.2 Do Video MLLMs Answer from the Moments They Attend To?

Existing rewards constrain the predicted interval, but nothing ties that interval to the evidence the model actually uses. We make this measurable by extracting the attended interval from prefill attention and comparing the two.

3.2.1 Extracting the Attended Interval

We extract the attended interval from prefill attention, following Du et al. (2026). We sample F frames from V; Qwen2.5-VL (Bai et al., 2025) merges adjacent frames with a temporal patch size of 2, so video tokens form $T = F/2$ temporal groups. A single prefill pass caches attention from the question-token rows to the video-token columns, retaining the head axis. For each calibrated head, we pool the question rows with weights $w_i = \exp(-H_i/\tau)$, where H_i is the entropy of row i and $\tau = 0.5$, so that rows committing to a few groups outweigh function words that spread attention over everything. This yields a temporal profile $a^+ \in \mathbb{R}^T$. Raw attention is dominated by position rather than content, so we repeat the pass with the video masked out under an identical prompt, obtaining a^-, and retain only the component specific to this video,

$$s = \max\left(a^+ \log \frac{a^+}{a^-},\ 0\right), \tag{3}$$

Table 1: Agreement between the attended and predicted intervals on Charades-STA. $v = \mathrm{IoU}(\mathcal{W}, \hat{\mathcal{W}})$ denotes the agreement between the interval extracted from attention and the interval predicted by the model, and c the event that the predicted interval attains IoU ≥ 0.5 against the annotated span. Accuracy is partitioned into grounded ($v \geq 0.5$) and ungrounded ($v < 0.5$), and the ungrounded rate is the ratio of the latter to accuracy.

Model	**Agree.** $v \geq 0.5 \uparrow$	**A-IoU** $\bar{v} \uparrow$	**Acc.** R@0.5 $\uparrow$	**Grounded** $\uparrow$	**Ungrounded** $\downarrow$	**Ungr. rate** $\downarrow$
Qwen2.5-VL-7B	0.212	0.289	0.371	0.132	0.239	0.644
GRPO	0.310	0.355	0.362	0.192	0.170	0.471
Video-R1-7B	0.315	0.338	0.326	0.177	0.149	0.458
VideoChat-R1	0.460	0.462	0.584	0.306	0.279	0.477
VideoAPR	**0.518**	**0.490**	**0.607**	**0.418**	**0.149**	**0.311**

where weighting the log-ratio by a^+ suppresses ratios that explode where attention mass is negligible. We then L_1-normalize, smooth with a Gaussian ($\sigma = 2.0$), and average over the calibrated heads. The attended interval $\mathcal{W}$ comprises the groups reaching $\rho = 0.8$ times the profile maximum, dilated by ± 2.5 s.

Calibration runs once, offline, on $1{,}000$ Charades-STA (Gao et al., 2017) training clips. Heads are ranked by the IoU between their extracted interval and the annotation, and the five with the highest cross-shard agreement over eight disjoint shards are retained. Heads are calibrated per model, so that every model in Section 3.2.2 is read with the heads that best expose its own temporal signal. Annotations are used only here and never enter the training loop. Without any training, the resulting extraction attains an mIoU of 0.480 and R@0.5 of 0.545 on Charades-STA, which is what makes it usable as a measurement instrument in the following section.

3.2.2 Metrics and Findings

We now use this extraction to ask whether RL-trained video MLLMs answer from the moments they attend to. Applying the procedure of Section 3.2.1 to each model with its own calibrated heads, let $\mathcal{W}$ be the interval extracted from its attention and $\hat{\mathcal{W}}$ the interval it predicts. We define the agreement $v = \mathrm{IoU}(\mathcal{W}, \hat{\mathcal{W}})$ and report its mean (A-IoU) together with the fraction of samples satisfying $v \geq 0.5$. Writing c for the event that $\hat{\mathcal{W}}$ attains IoU ≥ 0.5 against the annotated span, we partition accuracy into

$$\text{grounded} = \Pr\left[c \wedge v \geq 0.5\right], \qquad \text{ungrounded} = \Pr\left[c \wedge v < 0.5\right], \tag{4}$$

and report the ungrounded rate $\Pr\left[v < 0.5 \mid c\right]$. Accuracy alone conflates the two, since a prediction counted as correct may or may not be supported by the evidence the model used.

Table 1 reports these metrics for an untrained backbone and three RL-trained models. The backbone is correct largely by coincidence, with 64% of its correct predictions produced while its attention points elsewhere. RL reduces this rate to roughly 0.47 and then stalls there. VideoChat-R1 is the clearest case. It is trained with an explicit interval IoU reward and improves both agreement and accuracy by a wide margin, yet its ungrounded rate of 0.477 is indistinguishable from the 0.471 of vanilla GRPO. What the interval reward buys is more predictions of both kinds in the same proportion, not a shift in the proportion itself. Supervising the interval a model predicts therefore does not make that prediction follow from the evidence the model used, which is what the reward in Section 3.3 targets directly. Accuracy here is measured against the annotated span and is independent of the extraction, so the two axes of Table 1 do not share a source of error.

3.3 Attention-Proposed Reward

Building on this analysis, we reward a response for being recoverable from the interval the policy attends to. If $\mathcal{W}$ carries the evidence for a response, that response should remain likely once the rest of the video is removed, and should become less likely elsewhere in the clip.

We restrict the visual context to $\mathcal{W}$, giving V_{focus}, and rescore the response o_i the policy has already produced. An absolute score is not a sufficient test, since a response that survives an arbitrary crop scores well wherever $\mathcal{W}$ lies. We therefore contrast V_{focus} with a control window V_{contrast} of the same length, placed maximally far along the timeline, and take the length-normalized margin

$$\Delta_i = \frac{1}{|o_i|}\Big(\log p_\theta(o_i \mid q, V_{\text{focus}}) - \log p_\theta(o_i \mid q, V_{\text{contrast}})\Big). \tag{5}$$

Two properties follow from this form. The margin compares two conditions on the same response, so any competence that transfers across crops raises both terms and cancels, leaving only what is specific to $\mathcal{W}$. Because the control has the same length, enlarging $\mathcal{W}$ moves mass into both windows and does not increase the margin, which removes the need for the explicit length penalty that interval rewards require. Being continuous, the margin also remains informative when every window admits a correct answer, a regime in which a binary sufficiency test collapses to a constant and, by Equation 1, contributes no gradient.

Masking is applied to the attention mask rather than to the pixels, since blanking frames would present the vision tower with an input distribution it was never trained on. As M-RoPE derives token positions from the attention mask, positions are computed once from the unmasked layout and passed in, so that removing frames does not shift the timeline of those that remain.

Writing c_i for the event that o_i is correct under the unrestricted video, the total reward is

$$R_i = r_i^{\text{fmt}} + r_i^{\text{acc}} + \lambda\, \mathbf{1}[c_i]\, \Delta_i, \tag{6}$$

optimized with GRPO as in Section 3.1. The gate restricts credit to responses that are already correct, so that a response whose prediction breaks under cropping is not rewarded for fragility. Since the prompt is shared within a group, $\mathcal{W}$ is extracted once per optimization step and reused across all G rollouts, at the cost of two prefill passes. Scoring o_i under the two windows requires only teacher-forced forward passes, so no additional generation is incurred. Verification is confined to training, and at inference the model runs a single forward pass on the full video.

4 Experiments

4.1 Experimental Setup

Benchmark and metric. Following previous works (Feng et al., 2026; Wang et al., 2026a), we evaluate on six video reasoning and understanding benchmarks: VSI-Bench (Yang et al., 2025), VideoMMMU (Hu et al., 2026), MMVU (Zhao et al., 2025), MVBench (Li et al., 2024b), TempCompass (Liu et al., 2024) and VideoMME (Fu et al., 2025), covering spatial reasoning, knowledge-intensive video QA, temporal logic, and general video understanding. Following conventions, we use only the multiple-choice subset of MMVU, and VideoMME is evaluated without subtitles. Average accuracy is adopted as the evaluation metric.

Model training. We follow RFT to train VideoAPR in two stages, a warm-up SFT stage and an RL stage. For the SFT stage, which equips the model with the ability to produce structured reasoning traces for diverse questions, we adopt the checkpoint released by Video-R1 (Feng et al., 2026), fine-tuned from Qwen2.5-VL-7B (Bai et al., 2025) on Video-R1-CoT-165k. The RL stage is based on Video-R1-260k and uses the reward in Equation 6 to tie each response to the interval the policy attends to. Prior to both stages, the attention heads used for the extraction are calibrated once, offline, on Charades-STA (Gao et al., 2017) training clips as described in Section 3.2.1, and the resulting head set is reused unchanged, so that no temporal annotation enters the RL loop.

Implementation details. We train VideoAPR on 8 NVIDIA A100 GPUs with 80GB each, using full fine-tuning under DeepSpeed ZeRO-3. For efficiency, the video input is limited to 16 frames of resolution $128 \times 28 \times 28$ during training, where 28×28 is the size of each image patch and 128 denotes the number of patches. During inference, we increase the number of frames to 32 and the resolution to $256 \times 28 \times 28$. We sample $G = 8$ rollouts per prompt and train for 500 RL steps with a verification weight of $\lambda = 0.5$. The extraction uses the top five calibrated heads with a threshold of $\rho = 0.8$ and a dilation of 2.5 s, and the control window matches the length of the attended window.

Table 2: Performance comparison on six video reasoning and understanding benchmarks. During inference, the number of frames is 32 and the frame resolution is $256 \times 28 \times 28$. The best results are highlighted in bold.

Model	VSI.	VideoMMMU	MMVU	MV.	TempCom.	VideoMME
Proprietary Models						
GPT-4o	34.0	61.2	75.4	–	–	71.9
Open-Source Video MLLMs						
LLaMA-VID	–	–	–	41.9	45.6	–
ShareGPT4Video	–	–	–	51.2	–	39.9
VideoLLaMA2	–	–	44.8	54.6	–	47.9
LongVA-7B	29.2	23.9	–	–	56.9	52.6
VILA-1.5-8B	28.9	20.8	–	–	58.8	–
LLaVA-OneVision-7B	32.4	33.8	49.2	56.7	–	58.2
mPLUG-Owl3-8B	–	–	–	54.5	–	53.5
Qwen2.5-VL-7B	31.8	47.4	61.3	59.4	69.2	52.8
R1-based Video Reasoning Models						
TinyLLaVA-Video-R1	–	–	46.9	–	49.5	46.6
VideoChat-R1	–	–	–	67.9	–	–
Video-R1	35.8	**52.3**	63.8	63.9	73.2	59.3
VideoRFT	36.8	51.1	**68.5**	62.1	73.7	**59.8**
VideoAPR	**36.9**	**52.3**	65.9	**68.8**	**73.9**	59.4

4.2 Main Results

Quantitative comparison. Table 2 reports our comparison against proprietary models (Hurst et al., 2024), open-source video MLLMs (Li et al., 2024c; Chen et al., 2024; Cheng et al., 2024; Zhang et al., 2024; Lin et al., 2024; Li et al., 2024a; Ye et al., 2025; Bai et al., 2025), and recent R1-based video reasoning models (Zhang et al., 2025; Li et al., 2025; Feng et al., 2026; Wang et al., 2026a). Several observations can be drawn from the results. First, VideoAPR improves over its base model, Qwen2.5-VL-7B, on all six benchmarks, by $+9.4$ on MVBench, $+6.6$ on VideoMME and $+5.1$ on VSI-Bench, and outperforms every open-source MLLM on each benchmark for which results are reported. Second, on VSI-Bench it exceeds the proprietary GPT-4o by $+2.9$, suggesting that rewarding a policy for answering from the evidence it attends to narrows the gap to closed-source models on spatially and temporally grounded reasoning. Third, among recent RL-based models, VideoAPR ranks first on four of six benchmarks and attains the highest average accuracy at 59.5, against 58.7 for VideoRFT and 58.1 for Video-R1. The margins are widest on MVBench and TempCompass, the two benchmarks whose questions depend most directly on when an event occurs, which is where a reward defined over the attended interval should help. On MMVU and VideoMME, where scene-level content more often suffices, VideoAPR stays within 2.6 and 0.4 points of the best reported result.

Qualitative comparison. Figure 3 compares reasoning traces on a question from MVBench. Video-R1 enumerates the options at length and dismisses closing the window on the grounds that the person has been interacting with it, then concludes that the answer is A before emitting D, so its trace and its answer disagree. VideoChat-R1 is shorter but describes the person as reaching for something on the counter, which the frames do not show. Neither trace refers to the moment the question is about. VideoAPR describes the closing frames, where the person turns from the sink toward the window, and answers correctly. The two failure modes illustrate what an outcome reward leaves unconstrained, since a trace may be long and self-inconsistent, or short and fluent while describing content that is not in the video.



Q: What will the person do next?

(A) Take the food. (B) Close the door.
(C) Close the window. (D) Open the window.

Video-R1 ✗

Let me think about this for a moment. A. Take the food — this seems plausible because the person has been handling food. […] C. Close the window — this also doesn't seem likely because the person has been interacting with the window. […] Oh, I see, the most logical next step would be to take the food. […] Final answer: A. Let's double-check. Yes, the person will take the food next.

Answer: D

VideoChat-R1 ✗

The person has just finished washing dishes and is now reaching for something on the counter. Given the context, it seems logical that they might be taking a food item from the counter, which aligns with option A. The other options don't fit the immediate action of the person.

Answer: A

VideoAPR (ours) ✓

In the closing frames the person turns away from the sink and reaches up toward the open window with both hands, not toward the counter. The blinds are drawn back and the pane is still open at that point, so the action that follows is closing it.

Answer: C

Figure 3: Qualitative comparison of reasoning traces on MVBench. The correct answer is (C).

Table 3: Ablation on the verification reward. All variants share the same initialization and training budget.

Reward	VSI.	VideoMMMU	MMVU	MV.	TempCom.	VideoMME
GRPO	33.7	48.7	63.0	61.9	71.5	56.6
+ focus	35.6	50.1	63.4	64.1	72.7	57.5
+ contrast (ours)	**36.9**	**52.3**	**65.9**	**68.8**	**73.9**	**59.4**

4.3 Ablation Study

Table 3 isolates the contribution of each part of the verification reward. *GRPO* trains with the format and accuracy rewards only. *Focus* replaces Δ_i in Equation 6 with the first term of Equation 5, rewarding the likelihood of the response under the attended window alone. *Contrast* uses Δ_i as defined, subtracting the score under the control window. Both variants keep the correctness gate, and all three share the same initialization, data, and number of RL steps, so the differences are attributable to the reward alone.

Both components help, and the second helps more. The focus term raises the six-benchmark average by 1.3 points over GRPO, and the contrast term raises it by a further 2.3, so most of the total gain of 3.6 points comes from the comparison rather than from the attended window alone. This is consistent with the argument in Section 3.3. An absolute score rewards a response for surviving any crop, whereas the difference isolates what the attended window explains that another window does not. The pattern is sharpest on MVBench, where the focus term contributes 2.2 points and the contrast term a further 4.7. Gains are smaller on TempCompass, where GRPO already reaches 71.5 and the headroom is correspondingly limited.

Table 4: Performance comparison at a reduced temporal resolution. During inference, the number of frames is 16 and the frame resolution is $256 \times 28 \times 28$. The best results are highlighted in bold.

Model	VSI.	VideoMMMU	MMVU	MV.	TempCom.	VideoMME
Video-R1	34.6	**49.8**	64.2	62.7	72.6	57.4
VideoRFT	34.3	47.1	**65.3**	63.1	72.2	54.4
VideoAPR	**35.3**	48.8	64.1	**66.2**	**72.8**	**57.7**

4.4 Robustness to Temporal Resolution

We further compare VideoAPR with Video-R1 and VideoRFT under a reduced temporal resolution to verify that the effectiveness of our reward does not depend on a dense frame budget. This setting is particularly relevant to our method, since the extraction operates on $T = F/2$ temporal groups, so halving the number of sampled frames also halves the resolution at which an interval can be localized. At 32 frames the profile spans sixteen groups, whereas at 16 frames it spans only eight, and an interval thresholded at $\rho = 0.8$ consequently covers a larger fraction of the clip.

As shown in Table 4, VideoAPR still achieves the best performance on four out of six benchmarks, with the largest improvement of 3.1 points on MVBench. The relative ordering among the three models remains unchanged, demonstrating that our reward does not require a fine temporal grid to be effective. VideoAPR falls behind on VideoMMMU and MMVU, which are also the two benchmarks on which it gains least at 32 frames, indicating that the gap is not introduced by the reduced resolution.

5 Conclusion

We analyze whether reinforcement learning makes video MLLMs answer from the moments they attend to. Extracting the attended interval from prefill attention and comparing it against the interval each model predicts, we find that a large fraction of correct predictions are not supported by the model's own attention, and that this fraction does not improve with rewards that supervise the predicted interval. Outcome rewards constrain only the answer, and interval IoU constrains only what the model says about time, so neither requires the answer to follow from the evidence the model used.

Motivated by this, we propose VideoAPR, which rewards a response for being recoverable from the interval the policy attends to. The attended window is extracted from the model's own attention, and the reward is the margin by which that window explains the response better than an equal-length window elsewhere in the clip, granted only to responses that are already correct. The method requires no temporal annotation during training and leaves inference unchanged. Extensive experiments across six benchmarks demonstrate that VideoAPR outperforms existing methods and achieves state-of-the-art performance in video reasoning.

AI use statement

In this work, we used generative AI tools to aid and polish the writing of this manuscript. We have not used generative AI tools for the design or implementation of the proposed method, or for the execution of the experiments. We have reviewed all AI-assisted work: every AI-assisted passage was read and revised by the authors, and all reported numbers were produced by our own experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,

Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

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

Dazhao Du, Liao Duan, Jian Liu, Tao Han, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. Mllms know when before speaking: Revealing and recovering temporal grounding via attention cues. *arXiv preprint arXiv:2605.21954*, 2026.

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

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

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

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

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

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. Visioncoach: Reinforcing grounded video reasoning via visual-perception prompting. In *European Conference on Computer Vision*, pp. 620–637. Springer, 2026.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024b.

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 preprint arXiv:2504.06958*, 2025.

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

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

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

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044. IEEE, 2025.

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

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

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024.

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

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

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. In *International Conference on Learning Representations*, volume 2026, pp. 26868–26898, 2026.

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

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 98891–98913, 2025.

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

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

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

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025.

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

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