VEPO: LEARNING TO CREDIT VISUAL EVIDENCE SEEKING FOR AGENTIC VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video understanding requires acquiring sparse, temporally distributed visual evidence from long and redundant videos. Agentic approaches address this challenge through iterative temporal search, allowing models to gather visual observations before producing an answer. We find that correct answers produced by agentic systems are not always supported by the acquired evidence. During training with standard GRPO, answer accuracy improves while temporal evidence recall and active search decline, a training dynamic we characterize as *evidence-seeking collapse*. The objective rewards final-answer correctness, assigning identical rewards to evidence-supported predictions and unsupported guesses without checking whether sufficient evidence was acquired. Recent methods incorporate tool-use incentives and temporal localization rewards to encourage search activity and access to relevant regions, but these signals do not directly assess whether the accumulated observations provide sufficient evidence to answer the question. We introduce Video Evidence Policy Optimization (VEPO), which uses a frozen Evidence Reward Model to evaluate the sufficiency of accumulated observations. The resulting evidence signal allows VEPO to prefer evidence-supported answers over unsupported guesses and to credit the search decisions that improve evidence sufficiency. Across four long-video benchmarks, VEPO with a 7B backbone consistently outperforms supervised fine-tuning and standard reinforcement learning, reaching 63.0% on LongVideoBench and 47.5% on LVBench. These gains are accompanied by more correct answers supported by sufficient visual evidence, showing that VEPO improves accuracy while preserving grounded evidence acquisition. Our code is available at https://anonymous.4open.science/r/VEPO-0DD3/.

1 INTRODUCTION

Long-video understanding requires acquiring relevant visual evidence scattered across extensive temporal horizons (Xiao et al., 2021; Mangalam et al., 2023; Li et al., 2024a). Agentic approaches address this challenge by enabling models to actively acquire and refine visual evidence through iterative interaction with the video (Fan et al., 2024; Ma et al., 2025; Wang et al., 2025b). This explicit search provides the key advantage of agentic frameworks by making evidence acquisition explicit through an interaction trajectory (Yao et al., 2023). Training these agents therefore requires optimizing the interaction process that connects evidence acquisition with final prediction (Schick et al., 2023; Schulman et al., 2017).

However, exposing a search trajectory does not guarantee that the final prediction is actually grounded in the acquired visual evidence. Current training paradigms primarily optimize these trajectories through final-answer supervision, where correctness serves as the main learning signal (DeepSeek-AI, 2025). Although effective for improving answer accuracy, such objectives provide no guarantee that the acquired observations are sufficient to support the prediction. Under this objective, a trajectory that acquires the necessary visual support and an unsupported correct trajectory receive identical rewards. We characterize this failure mode as *evidence-seeking collapse*, a training dynamic where answer accuracy improves without preserving evidence acquisition. The policy maximizes its expected return by exploiting parametric priors and dataset shortcuts, bypassing the need to actively resolve visual uncertainty.

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.

This collapse originates from an unresolved ambiguity between outcome correctness and semantic evidence sufficiency. A correct answer alone does not reveal whether the accumulated observations jointly contain enough visual information to justify the prediction (Xiao et al., 2024). Recent methods attempt to encourage interaction through tool-use incentives and temporal localization rewards, but these signals evaluate where or how the agent searches rather than whether the accumulated observations are sufficient to support the answer. An agent can visit the relevant temporal region and still answer correctly without actually acquiring the necessary visual detail. Conversely, an agent can efficiently gather sparse observations that jointly resolve the query, yet receive low temporal rewards for failing to cover a continuous annotated interval. High temporal overlap neither guarantees nor is strictly required for resolving the query, so these localization signals do not reliably distinguish grounded evidence acquisition from unsupported guesses.

This distinction motivates an evidence-aware optimization objective that couples final correctness with evidence sufficiency and explicitly prefers grounded success over unsupported correctness. We introduce Video Evidence Policy Optimization (VEPO), which incorporates evidence sufficiency into reinforcement learning through a frozen Evidence Reward Model (ERM). The ERM evaluates whether the accumulated observations along a trajectory provide sufficient support for the final answer, allowing VEPO to distinguish grounded successes from unsupported guesses. Beyond trajectory-level assessment, VEPO assigns direct credit to the specific search actions that induce a transition toward evidence sufficiency, aligning local exploration behavior with grounded outcomes. VEPO consistently improves answer accuracy across four long-video benchmarks while strengthening the visual support for its predictions.

We make the following contributions:

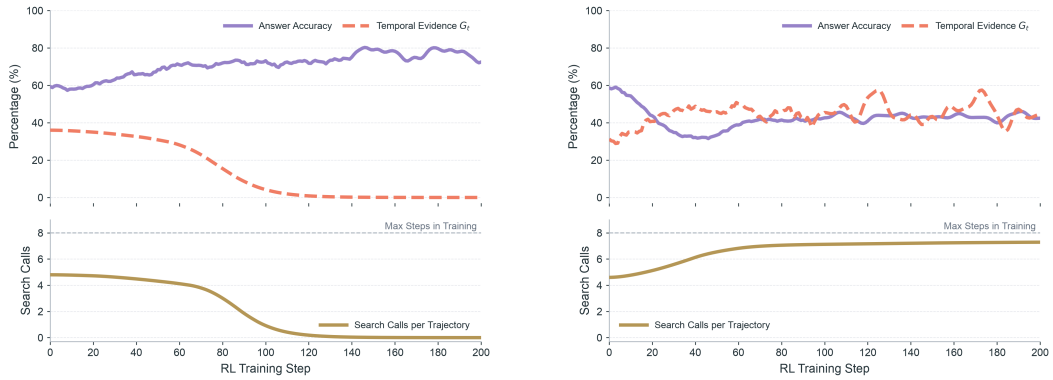
1. We formalize *evidence-seeking collapse* in agentic video reinforcement learning, a training dynamic where policy optimization improves answer accuracy while actively degrading temporal search. We demonstrate that this failure originates from an unresolved ambiguity in outcome-only rewards, which cannot distinguish genuinely grounded reasoning from unsupported correct guesses.
2. We introduce VEPO, a framework that resolves this outcome–evidence ambiguity by establishing semantic evidence sufficiency as an explicit optimization target. By evaluating the joint visual support accumulated along an interaction trajectory, VEPO translates this evidential judgment directly into trajectory-level preferences and action-local credit.
3. We demonstrate that aligning reinforcement learning with evidence sufficiency successfully reverses evidence-seeking collapse. VEPO consistently improves long-video reasoning performance while sustaining grounded visual interaction, ensuring that correct predictions are actively supported by acquired visual evidence.

2 Diagnosing Evidence-Seeking Collapse

In this section, we diagnose the phenomenon of evidence-seeking collapse in agentic video understanding. We characterize evidence acquisition in Section 2.1. We then examine training dynamics under outcome-only reinforcement learning and assess whether correct answers from existing agents are supported by sufficient retrieved evidence under current search incentives in Section 2.2. These analyses illustrate the unresolved ambiguity between outcome correctness and visual evidence sufficiency.

2.1 Characterizing Evidence Acquisition

Given a video $\mathcal{V}$ and a query q, a multimodal policy π_θ first receives a coarse global overview o_0 and interacts with the environment for at most K rounds. At round t, the policy conditions on the history h_{t-1} to generate a textual rationale z_t and an action u_t. This action specifies either a temporal window $w_t = [s_t, e_t]$ to inspect or a terminal answer $\hat{a}$. A window action returns the raw video frames within that interval as the visual observation o_t, updating the history to $h_t = h_{t-1} \oplus (z_t, u_t, o_t)$ without any intermediate reward. An answer action provides no further observation and terminates the rollout at round $T \leq K$, producing the complete interaction trace $\xi = h_T$.



(a) Outcome-only GRPO **(b) GRPO + Tool Reward**

Figure 1: Evidence-seeking collapse under outcome-only RL and the limits of rewarding tool use as a remedy. (a) Under direct GRPO, training answer accuracy improves while temporal evidence recall and search interaction collapse. (b) A generic tool-use reward drives search calls toward the interaction limit without a corresponding gain in answer accuracy.

Evaluating this search process requires distinguishing the final prediction from the visual evidence acquired to support it. We define $C(\xi) \in \{0, 1\}$ to indicate final-answer correctness, yielding the overall answer accuracy $\mathbb{E}[C(\xi)]$.

Temporal evidence. To quantify evidence acquisition independent of this final answer, we measure whether the sequence of retrieved windows $\mathcal{W}(\xi)$ reaches the ground-truth evidence intervals $\mathcal{G}$. The temporal grounding indicator

$$G_t(\xi) = \max_{w \in \mathcal{W}(\xi),\, g \in \mathcal{G}} \left(\mathrm{tIoU}(w, g) \geq \gamma\right), \tag{1}$$

captures whether at least one search window overlaps with the annotated evidence beyond a threshold γ. The resulting temporal evidence recall $\mathbb{E}[G_t(\xi)]$ isolates the policy's ability to locate relevant information. A trajectory satisfying $C(\xi) = G_t(\xi) = 1$ exhibits temporally grounded correctness, confirming that the agent found the required temporal region before answering correctly.

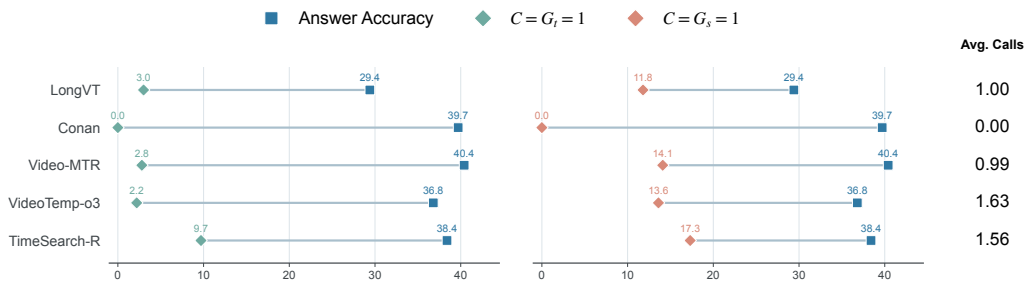
Semantic evidence. Evaluating the success of an interaction trajectory requires assessing whether the retrieved observations $\mathcal{O}(\xi) = \{o_t\}_{1 \leq t < T}$ jointly supply the visual information necessary to resolve the query. We define the binary semantic sufficiency indicator

$$G_s(\xi) = J(q, \mathcal{O}(\xi)). \tag{2}$$

where J assesses whether the accumulated observations jointly provide sufficient evidence. We implement J using an LLM-as-a-judge approach in the main evaluations. This diagnostic separates *supported correctness* ($C(\xi) = G_s(\xi) = 1$) from *unsupported success* ($C(\xi) = 1, G_s(\xi) = 0$), where the policy predicts the right answer without acquiring sufficient visual information. Such predictions may rely on parametric priors or dataset shortcuts.

2.2 Evidence-Seeking Collapse and Unsupported Success

The prevailing paradigm for training agentic video systems relies on final-answer correctness to optimize the multi-turn interaction trajectory. To examine the effect of this outcome-only supervision, we train this long-video agent on CG-Bench (Chen et al., 2025a) with direct GRPO using only final-answer rewards. We observe that answer accuracy rises while temporal evidence recall and active search calls continuously decline in Figure 1a. This divergence reveals *evidence-seeking collapse*: optimization improves the terminal answer predictor while degrading the temporal search policy. Because the learning objective rewards $C(\xi)$ without requiring $G_t(\xi) = 1$, the policy maximizes its return by exploiting parametric priors and the initial overview, leaving visual uncertainty unresolved. The rising answer accuracy therefore systematically masks a loss of grounded interaction.



(a) Temporal Evidence **(b)** Semantic Evidence

Figure 2: Cross-method evidence diagnostics on LVBench. Dumbbells illustrate the gap between answer accuracy and temporally grounded correctness ($C = G_t = 1$) in (a), and supported correctness ($C = G_s = 1$) when the frozen evidence judge jointly evaluates all observations retrieved along each trajectory in (b). The rightmost column reports average tool calls per query.

Addressing this collapse requires encouraging evidence acquisition, yet standard interaction incentives remain insufficient. Supplementing the outcome-only objective with a per-call tool reward drives search calls toward the interaction limit without a corresponding improvement in answer accuracy in Figure 1b. Rewarding execution frequency can encourage redundant temporal queries without ensuring meaningful evidence acquisition.

To guide temporal search, recent frameworks incorporate temporal localization rewards to assess overlap with annotated evidence regions (Yang et al., 2025; Liu et al., 2026). We examine the predictions of five recent agentic frameworks on LVBench (Yang et al., 2025; Ouyang et al., 2025; Xie et al., 2025; Liu et al., 2026; Pan et al., 2025), including those trained with localization signals. Across all evaluated methods, answer accuracy substantially exceeds both temporally grounded correctness in Figure 2a and semantically supported correctness in Figure 2b. This outcome–evidence gap is accompanied by low average search counts across these baselines. The limited search activity suggests that agents often terminate search before acquiring sufficient visual evidence and rely on guessing to produce an answer. This gap remains visible among policies trained with temporal localization rewards, revealing that visiting a relevant temporal region does not guarantee that the accumulated frames jointly contain the visual details required to answer the query. These localization signals evaluate the temporal location of the search but do not directly assess the semantic sufficiency of the acquired visual content. These findings motivate an objective that couples final-answer correctness with the joint semantic sufficiency of the acquired visual evidence.

3 METHOD

In this section, we introduce Video Evidence Policy Optimization (VEPO) to address evidence-seeking collapse by aligning outcome correctness with visual evidence acquisition. Our presentation proceeds in three steps. First, we provide an overview of the VEPO framework and its policy optimization procedure in Section 3.1. Then, we detail the Evidence Reward Model used to assess joint semantic sufficiency across accumulated observations in Section 3.2. Finally, we introduce how to translate this continuous evidence judgment into a trajectory-level preference and an action-local credit in Section 3.3. Together, these steps connect the final prediction to the specific search actions that acquire the supporting evidence.

3.1 VEPO FRAMEWORK

VEPO addresses evidence-seeking collapse by making evidence sufficiency explicit in policy optimization, as illustrated in Figure 3. A frozen ERM evaluates accumulated observations at each trajectory prefix, providing a shared signal for trajectory-level preference and action-local evidence credit. Final sufficiency and answer correctness define the trajectory reward, from which GRPO computes group-relative advantages after dynamic group filtering. Positive sufficiency-score changes provide additional credit to the corresponding search-action tokens in correct trajectories.

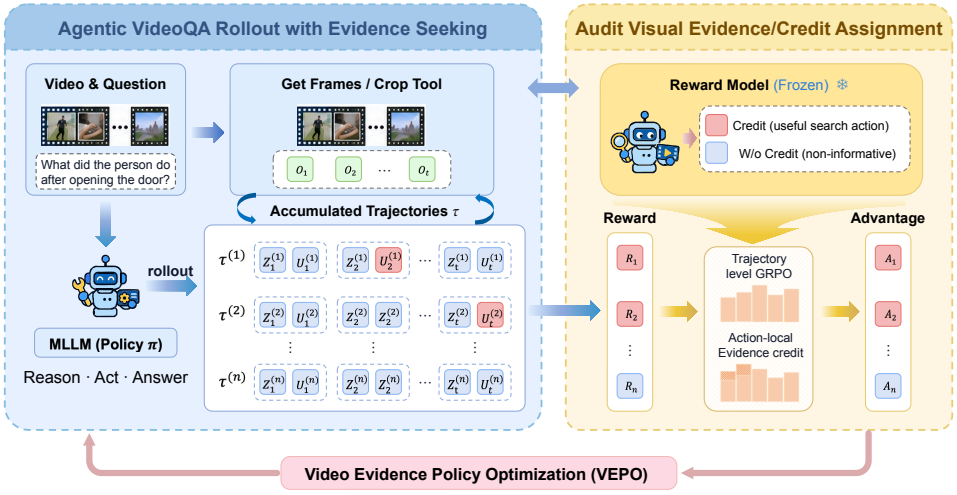


Figure 3: Overview of the VEPO framework.

3.2 Assessing Joint Evidence Sufficiency

Evaluating evidence acquisition requires assessing whether the accumulated observations jointly resolve the query. Answer-relevant events are often temporally distributed, so individual search windows may capture only partial clues. These scattered observations become sufficient only when combined, requiring an independent judge to evaluate their complementary information. We introduce an Evidence Reward Model (ERM) to provide this continuous sufficiency judgment. For an interaction prefix ending at step t, the ERM takes the question q and the complete set of retrieved visual observations $\mathcal{O}_t = \{o_j\}_{1 \le j \le t}$ to estimate the probability of joint sufficiency, yielding the prefix score $s_\phi(\xi_{\le t}) = p_\phi(y_t = 1 \mid q, \mathcal{O}_t)$, where $y_t \in \{0, 1\}$ indicates whether the accumulated context contains enough evidence to support a correct prediction.

To train the ERM to evaluate joint reasoning across windows, we use synthesized multi-window contexts that mimic the evolving visual state of an active search trajectory. Positive contexts contain the complete set of supporting windows alongside distractor observations. Negative contexts preserve the identical distractor composition but omit necessary evidence, ensuring that queries requiring multiple temporal events remain classified as insufficient until the final supporting window is appended. This construction forces the ERM to recognize the exact transition where an incomplete visual history becomes sufficient.

During subsequent policy optimization, the ERM acts as a frozen evidence judge. By applying this joint assessment to the observations accumulated at each trajectory prefix, the ERM generates a sequence of sufficiency scores $\{s_\phi(\xi_{\le t})\}_{t<T}$. This continuous signal captures how each active search alters the evidential state, providing the necessary foundation for crediting evidence-seeking actions. Data construction and training details are provided in Appendices A.1 and B.

3.3 Trajectory Preference and Action-Local Credit

Trajectory-level preference. To resolve the outcome-evidence ambiguity identified in Section 2.2, VEPO first breaks the reward equivalence between supported and unsupported success. For each query, we sample a group of complete interaction trajectories and evaluate both final-answer correctness $C(\xi)$ and the terminal evidence sufficiency score $s_\phi(\xi)$. We use the final prefix score $s_\phi(\xi) = s_\phi(\xi_{\le T-1})$ to assess trajectory-level evidence sufficiency, with τ denoting the sufficiency threshold. We structure the trajectory reward $R_{\text{traj}}(\xi)$ to enforce a strict preference hierarchy:

$$R_{\text{traj}}(\xi) = \begin{cases} r_{\text{wrong}}, & C(\xi) = 0, \\ r_{\text{corr}}, & C(\xi) = 1,\ s_\phi(\xi) < \tau, \\ r_{\text{evi}}, & C(\xi) = 1,\ s_\phi(\xi) \ge \tau. \end{cases} \tag{3}$$

Table 1: Accuracy on four long-video understanding benchmarks. All agentic baselines use 64 overview frames and at most 16 frames per call, without external models for fair comparison. † indicates removal of CLIP retriever.

Models	Backbone	MLVU	VideoMME (w/o sub)		LongVideoBench	LVBench
			Overall	Long		
Duration	–	3∼120 min	1∼60 min	30∼60 min	0∼60 min	4101 sec
Proprietary Models						
Gemini-1.5-Pro	Gemini-1.5-Pro	61.8	75.0	67.4	64.0	33.1
GPT-4o	GPT-4o	66.2	77.2	72.1	66.7	34.7
Open Source Video MLLMs						
Video-LLaVA (Lin et al., 2024)	Vicuna-7B-v1.5	47.3	40.4	38.1	39.1	-
LongVA (Zhang et al., 2025c)	Qwen2-7B	56.3	54.3	47.6	-	-
Video-R1 (Feng et al., 2025)	Qwen2.5-VL-7B	45.4	59.3	-	-	35.9
Qwen2.5-VL-Instruct-7B (Bai et al., 2025)	Qwen2.5-VL-7B	63.9	58.4	44.9	55.3	34.6
End to End Agentic Frameworks						
VideoAgent (Wang et al., 2024a)	GPT-4o	55.8	59.4	47.9	50.3	42.3
Conan (Ouyang et al., 2025)	Qwen2.5-VL-7B	63.6	60.8	48.4	57.1	39.7
Video-MTR (Xie et al., 2025)	Qwen2.5-VL-7B	55.8	60.0	48.2	54.2	40.4
LongVT (Yang et al., 2025)	Qwen2.5-VL-7B	58.0	58.1	49.7	41.6	29.4
VideoTemp-o3 (Liu et al., 2026)	Qwen2.5-VL-7B	59.8	58.3	48.4	53.2	36.8
TimeSearch-R† (Pan et al., 2025)	Qwen2.5-VL-7B	67.1	62.8	50.2	48.4	38.4
SFT	Qwen2.5-VL-7B	64.7	61.8	49.7	58.3	43.8
GRPO	Qwen2.5-VL-7B	63.2	60.5	48.1	56.9	42.3
GRPO + Tool Reward	Qwen2.5-VL-7B	60.9	58.7	48.7	57.8	43.8
VEPO (Ours)	Qwen2.5-VL-7B	**68.0**	**64.4**	**51.4**	**63.0**	**47.5**

Table 2: Cross-method evidence diagnostics on LVBench using the temporal and semantic evidence metrics defined in Section 2.1. Temporal evidence is evaluated on questions with temporal annotations using $\gamma = 0.05$ (Chen et al., 2025a; Qiu et al., 2026).

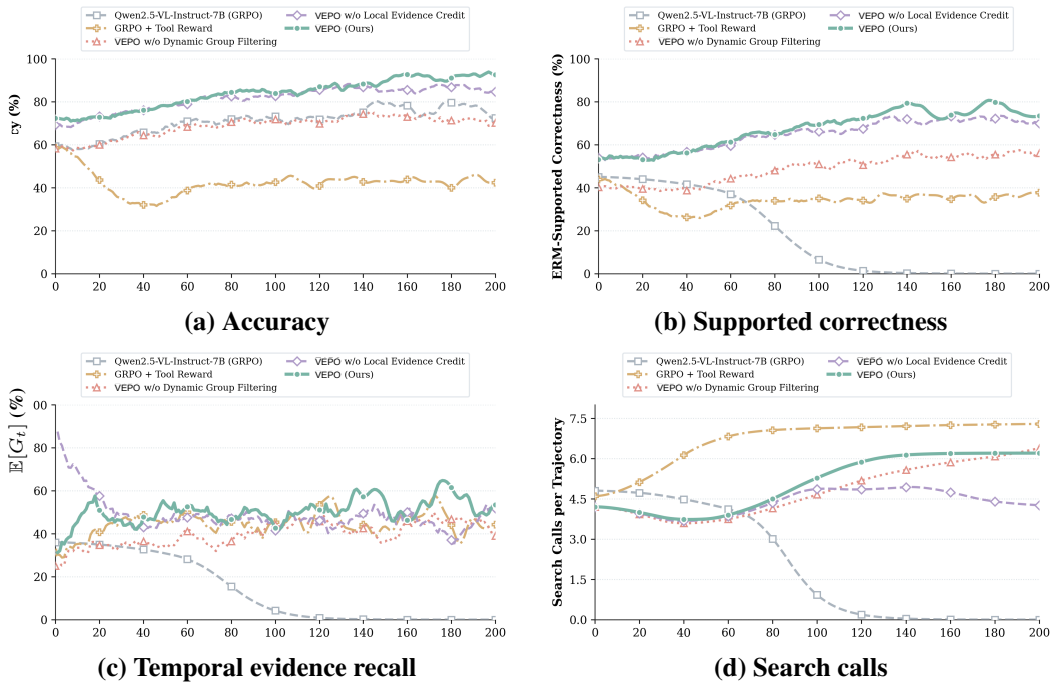
Method	Acc.	Temporal Evidence		Semantic Evidence		Avg. Calls
		$G_t = 1$	$C = G_t = 1$	$G_s = 1$	$C = G_s = 1$	
LongVT	29.4	6.4	3.0	34.3	11.8	1.00
Conan	39.7	0.0	0.0	0.0	0.0	0.00
Video-MTR	40.4	5.1	2.8	30.0	14.1	0.99
VideoTemp-o3	36.8	5.4	2.2	32.6	13.6	1.63
TimeSearch-R	38.4	20.0	9.7	40.2	17.3	1.56
VEPO (Ours)	**47.5**	**39.1**	**21.5**	**66.5**	**36.3**	**3.18**

By setting $r_{\text{wrong}} < r_{\text{corr}} < r_{\text{evi}}$, the objective explicitly prefers an evidence-supported correct trajectory over an unsupported guess. GRPO (Shao et al., 2024) translates these rewards into a group-relative trajectory advantage A_{traj}. To ensure stable optimization, we concentrate policy updates on rollout groups that contain an evidence-supported trajectory and exhibit meaningful reward variation, providing the optimizer with a clear contrast between grounded and ungrounded outcomes.

Action-local evidence credit. While A_{traj} successfully steers the policy toward the preferred outcome, this global advantage is uniformly distributed across the entire interaction sequence. It rewards the trajectory as a whole but fails to distinguish the specific search action that successfully acquired the missing evidence from redundant or irrelevant calls earlier in the rollout. To resolve this credit assignment problem, we extract an action-local evidence credit from the continuous state changes evaluated by the ERM. For a correct trajectory, the token span generating the search action at step t receives the augmented advantage:

$$A_t = A_{\text{traj}} + \max\{0,\ s_\phi(\xi_{\le t}) - s_\phi(\xi_{\le t-1})\}. \tag{4}$$

This formulation compares the accumulated visual evidence before and after the temporal search. The positive difference directly credits the specific action that induces a transition toward sufficiency, providing additional action-local credit where the evidence state improves. All remaining tokens, including the final answer rationale, retain the baseline trajectory advantage A_{traj}.



(a) Accuracy **(b) Supported correctness**

(c) Temporal evidence recall **(d) Search calls**

Figure 4: Training dynamics under a shared SFT model and RL data. Direct GRPO optimizes final-answer correctness, whereas GRPO + Tool Reward additionally rewards tool use. Supported correctness is the fraction of trajectories that answer correctly and whose accumulated multi-window observations are jointly accepted by the frozen ERM.

4 Experiments

In this section, we present the main evaluation on answer accuracy and evidence support across four long-video benchmarks in Section 4.1, as well as how VEPO reverses evidence-seeking collapse during training in Section 4.2. Finally, we describe the ablation studies on dynamic group filtering and local evidence credit in Section 4.3, followed by an analysis of observation and search budgets in Section 4.4.

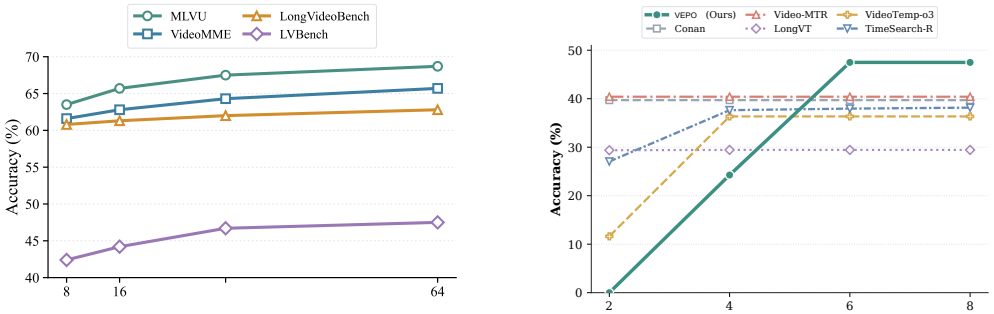
4.1 Answer Accuracy and Evidence Support

Answer accuracy. As shown in Table 1, VEPO consistently improves long-video question answering across all four benchmarks, outperforming the compared agentic frameworks and open-source video MLLMs. The controlled comparison under a fixed Qwen2.5-VL-7B backbone and SFT initialization isolates the effect of the optimization objective. Direct GRPO degrades the cold-start SFT performance across all four benchmarks, and adding a tool-use reward fails to improve upon SFT. VEPO instead yields consistent improvements, lifting LongVideoBench accuracy from 58.3% to 63.0%. This divergence shows that generic outcome-only post-training and tool-use incentives are insufficient in this setting, while an objective aligned with evidence acquisition yields sustained task improvements.

Evidence-supported correctness. To examine whether these accuracy gains are accompanied by stronger visual evidence, we decompose the final predictions according to their evidential support in Table 2, using the frozen training-time ERM to jointly judge all observations retrieved along each trajectory. VEPO more than doubles the rate of supported correctness ($C = G_s = 1$) compared to the strongest baseline, from 17.3% to 36.3%. This semantic improvement is accompanied by a similar gain in temporally grounded correctness ($C = G_t = 1$), from 9.7% to 21.5%. Together, these gains show that more correct answers are supported by relevant and jointly sufficient visual evidence, narrowing the outcome–evidence gap identified in Section 2.2. The improvement

Table 3: Ablation of dynamic group filtering and local evidence credit on four LVU benchmarks.

Method	MLVU	VideoMME	LongVideoBench	LVBench
Qwen2.5-VL-Instruct-7B (GRPO)	63.2	60.5	56.9	42.3
Qwen2.5-VL-Instruct-7B (GRPO + Tool Reward)	60.9	58.7	57.8	43.8
VEPO w/o Dynamic Group Filtering	66.5	63.7	61.9	46.6
VEPO w/o Local Evidence Credit	65.4	61.9	61.7	45.1
VEPO	**68.0**	**64.4**	**63.0**	**47.5**



(a) Overview Frames **(b) Search Horizon**

Figure 5: Scaling evidence acquisition. (a) Accuracy across four LVU benchmarks as the initial full-video overview grows from 8 to 64 frames. (b) Accuracy on LVBench under increasing search horizons.

in evidence-supported correctness is also observed under an independent model judge and blinded human assessment, as reported in Appendix C.4.

4.2 Reversing Evidence-Seeking Collapse

We track the training dynamics of different reward objectives to assess whether optimizing for evidence sufficiency mitigates evidence-seeking collapse. The evaluation contrasts outcome-only supervision through direct GRPO and generic interaction incentives through a tool-use reward against our evidence-aware formulation, VEPO. Monitoring answer accuracy, search volume, and evidence support over training reveals whether improvements in answer prediction are accompanied by sustained evidence acquisition.

In Figure 4, direct GRPO improves training answer accuracy while temporal evidence recall and active search calls decline over training. Supplementing this outcome-only objective with a generic tool-use reward increases interaction volume without a corresponding improvement in supported correctness. Increased execution frequency therefore does not ensure sufficient visual evidence. VEPO couples accuracy improvements with sustained temporal recall and a rising rate of evidence-supported correctness. Removing either dynamic group filtering or local evidence credit weakens this coupled improvement, indicating that both trajectory-level contrast and action-level attribution contribute to sustaining evidence acquisition during optimization. These training dynamics show that VEPO mitigates evidence-seeking collapse while improving answer accuracy and supported correctness.

4.3 Understanding the Role of Evidence Alignment

We ablate dynamic group filtering and local evidence credit to examine how trajectory-level contrast and action-level attribution contribute to evidence acquisition. Both variants retain the same evidence-aware trajectory reward. In Table 3, removing local evidence credit causes a larger accuracy drop across all benchmarks than removing group filtering, indicating that trajectory-level rewards alone provide insufficient guidance for individual search decisions. Without local attribution, the trajectory advantage does not distinguish search actions that improve evidence sufficiency from calls that contribute little additional information. The performance reduction without group filtering further supports concentrating updates on rollouts with meaningful outcome–evidence contrast.

4.4 Scaling Evidence Acquisition

Performance in agentic video understanding relies on both the initial global context and the agent's capacity for subsequent temporal inspection. Increasing the density of the uniformly sampled overview yields monotonic accuracy gains across all four benchmarks in Figure 5a. This consistent improvement demonstrates that a richer global prior establishes a stronger foundation for the policy to identify which temporal regions require targeted search.

While a dense overview anchors the interaction, resolving specific visual queries ultimately requires active temporal inspection. Holding the initial overview fixed, we examine the interaction horizon required to resolve this remaining uncertainty. As the maximum number of allowed search actions increases on LVBench, VEPO's accuracy improves but saturates after six searches in Figure 5b. Additional interaction budget beyond this point provides no measurable gain. This plateau indicates that increasing the search horizon beyond six actions does not improve accuracy under the evaluated observation budget.

5 Related Work

MLLMs for Long Video Understanding. Recent video MLLMs extend temporal coverage through longer language contexts (Chen et al., 2025c; Shen et al., 2025b), denser frame sampling (Wang et al., 2024b; Cheng et al., 2024), hierarchical token compression (Li et al., 2024b; Xu et al., 2024; Shen et al., 2025a), and long-video instruction tuning (Li et al., 2025a; Zhang et al., 2023; Maaz et al., 2024; Zhang et al., 2025c; Li et al., 2025b; Zhang et al., 2025d; Chen et al., 2025b). These advances encode longer videos in a single forward pass (Song et al., 2024; Wang et al., 2025a; Ren et al., 2024), but fixed sampling or compression still determines the available evidence. Broader context improves coverage without identifying regions that need closer inspection, motivating adaptive visual computation during reasoning.

Agentic Evidence Seeking and Reinforcement Learning. Agentic systems replace a predetermined visual context with iterative access to the video. Early approaches coordinate planners, retrievers, grounders, and verifiers to identify relevant frames or temporal segments at inference time (Yu et al., 2023; Huang et al., 2024; Wang et al., 2024a; Liu et al., 2025). Recent frameworks learn temporal cropping or frame selection policies through supervised and reinforcement learning (Yang et al., 2025; Zhang et al., 2025a; He et al., 2025; Ouyang et al., 2025; Xie et al., 2025; Zeng et al., 2026; Liu et al., 2026; Yang et al., 2026). Their objectives combine verifiable answer rewards with temporal overlap, frame relevance (Lei et al., 2021), tool validity, localization, or exploration efficiency. These signals encourage temporal interaction but do not reveal whether a correct trajectory has acquired sufficient visual support for its answer. Trajectory-level outcome signals evaluate complete rollouts while leaving their evidential basis unresolved (Uesato et al., 2022; Lightman et al., 2024). Local relevance signals evaluate individual observations without establishing whether the accumulated evidence supports the final answer. VEPO connects these two levels by distinguishing supported from unsupported success over complete trajectories and attributing improvements in evidence sufficiency to the search actions that acquire support (Setlur et al., 2025).

6 Conclusion

We identify evidence-seeking collapse in agentic long-video reinforcement learning, where answer accuracy can improve while evidence acquisition deteriorates. Because outcome-only rewards fail to distinguish answers grounded in retrieved visual evidence from unsupported guesses, the policy learns to abandon active temporal search even as its apparent accuracy improves. To resolve this outcome-evidence ambiguity, we introduce VEPO, a framework that establishes evidence-supported correctness as the explicit target for policy optimization. VEPO jointly assesses the accumulated visual context and attributes evidence gains directly to the temporal search actions that produce them. Our results demonstrate that this alignment successfully reverses the collapse, yielding policies that consistently improve reasoning accuracy while sustaining grounded visual interaction. Ultimately, these findings establish that scalable agentic reasoning requires learning objectives that reward correct answers together with the targeted acquisition of the visual evidence required to support them.

Ethics Statement

Our research aims to improve evidence-supported reasoning in agentic video understanding for research and beneficial applications. We use publicly available datasets and pretrained models, and describe our data preparation, training procedures, and evaluation protocols to support transparency and reproducibility. We recognize that video datasets and language models may reflect biases in their source data and encourage responsible downstream use, attention to privacy, and continued research into fairness. We support the use of these resources in accordance with applicable licenses and terms of service.

Reproducibility Statement

We describe data construction in Appendix A, training hyperparameters in Appendix B, and the interaction and evidence-judge prompts in Appendix F. The evaluation protocols include independent assessments of evidence support in Appendix C.4. Upon acceptance, we will release our code, model weights, and data preparation scripts, together with training data where redistribution is permitted.

Use of LLMs

We used LLMs to generate multi-round search-and-answer trajectories for cold-start supervised fine-tuning. The generated trajectories were filtered for answer correctness and interaction consistency, as described in Appendix A. We also used LLMs to polish the grammar and phrasing of this manuscript. The authors reviewed and edited the writing suggestions and take responsibility for the training data and the final content of the paper.

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, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=le4IoZZHy1.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/22a7476e4fd36818777c47e666f61a41-Abstract-Datasets_and_Benchmarks_Track.html.

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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5,*

2025, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/fc2865a2fdd5fd401050d37ee4946850-Abstract-Conference.html.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025c. URL https://openreview.net/forum?id=wCXAlfvCy6.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024. doi: 10.48550/ARXIV.2406.07476. URL https://doi.org/10.48550/arXiv.2406.07476.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. videoagent: A memory-augmented multimodal agent for video understanding. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXII*, volume 15080 of *Lecture Notes in Computer Science*, pp. 75–92. Springer, 2024. doi: 10.1007/978-3-031-72670-5_5. URL https://doi.org/10.1007/978-3-031-72670-5_5.

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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/8eb3976840e08b80dda9667562574246-Abstract-Conference.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *CoRR*, abs/2509.24304, 2025. doi: 10.48550/ARXIV.2509.24304. URL https://doi.org/10.48550/arXiv.2509.24304.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: language instructed temporal-localization assistant. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV*, volume 15122 of *Lecture Notes in Computer Science*, pp. 202–218. Springer, 2024. doi: 10.1007/978-3-031-73039-9_12. URL https://doi.org/10.1007/978-3-031-73039-9_12.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *CoRR*, abs/2107.09609, 2021. URL https://arxiv.org/abs/2107.09609.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22195–22206. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: chat-centric video understanding. *Sci. China Inf. Sci.*, 68(10), 2025a. doi: 10.1007/S11432-024-4321-9. URL https://doi.org/10.1007/s11432-024-4321-9.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *CoRR*, abs/2501.00574, 2025b. doi: 10.48550/ARXIV.2501.00574. URL https://doi.org/10.48550/arXiv.2501.00574.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, volume 15104 of *Lecture Notes in Computer Science*, pp. 323–340. Springer, 2024b. doi: 10.1007/978-3-031-72952-2_19. URL https://doi.org/10.1007/978-3-031-72952-2_19.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.342. URL https://doi.org/10.18653/v1/2024.emnlp-main.342.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Unleashing hour-scale video training for long video-language understanding. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/19741c617c87a25627682e5714af1501-Abstract-Conference.html.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, Yinwei Wei, and Xuemeng Song. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *CoRR*, abs/2602.07801, 2026. doi: 10.48550/ARXIV.2602.07801. URL https://doi.org/10.48550/arXiv.2602.07801.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *CoRR*, abs/2503.13444, 2025. doi: 10.48550/ARXIV.2503.13444. URL https://doi.org/10.48550/arXiv.2503.13444.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18936–18946. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01764. URL https://openaccess.thecvf.com/

content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.679. URL https://doi.org/10.18653/v1/2024.acl-long.679.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *CoRR*, abs/2510.20470, 2025. doi: 10.48550/ARXIV.2510.20470. URL https://doi.org/10.48550/arXiv.2510.20470.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *CoRR*, abs/2511.05489, 2025. doi: 10.48550/ARXIV.2511.05489. URL https://doi.org/10.48550/arXiv.2511.05489.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. Videoseal: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. *CoRR*, abs/2605.12571, 2026. doi: 10.48550/ARXIV.2605.12571. URL https://doi.org/10.48550/arXiv.2605.12571.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14313–14323. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01357. URL https://doi.org/10.1109/CVPR52733.2024.01357.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open

language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025a. URL https://proceedings.mlr.press/v267/shen25j.html.

Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Yifan Zhang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiawu Zheng, Yan Zhang, Yiyi Zhou, Ran He, Caifeng Shan, Rongrong Ji, and Xing Sun. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuracy. *CoRR*, abs/2502.05177, 2025b. doi: 10.48550/ARXIV.2502.05177. URL https://doi.org/10.48550/arXiv.2502.05177.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18221–18232. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01725. URL https://doi.org/10.1109/CVPR52733.2024.01725.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025a. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025b. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *CoRR*, abs/2211.14275, 2022. doi: 10.48550/ARXIV.2211.14275. URL https://doi.org/10.48550/arXiv.2211.14275.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pp. 58–76. Springer, 2024a. doi: 10.1007/978-3-031-72989-8_4. URL https://doi.org/10.1007/978-3-031-72989-8_4.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pp. 396–416. Springer, 2024b. doi: 10.1007/978-3-031-73013-9_23. URL https://doi.org/10.1007/978-3-031-73013-9_23.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 24170–24181. IEEE, 2025a. doi: 10.1109/ICCV51701.2025.02240. URL https://doi.org/10.1109/ICCV51701.2025.02240.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for LLM

reasoning on long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 3272–3283. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.00311. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VideoTree_Adaptive_Tree-based_Video_Representation_for_LLM_Reasoning_on_Long_CVPR_2025_paper.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9777–9786. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00965. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13204–13214. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01254. URL https://doi.org/10.1109/CVPR52733.2024.01254.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *CoRR*, abs/2508.20478, 2025. doi: 10.48550/ARXIV.2508.20478. URL https://doi.org/10.48550/arXiv.2508.20478.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See-Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *CoRR*, abs/2404.16994, 2024. doi: 10.48550/ARXIV.2404.16994. URL https://doi.org/10.48550/arXiv.2404.16994.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. Longvt: Incentivizing "thinking with long videos" via native tool calling. *CoRR*, abs/2511.20785, 2025. doi: 10.48550/ARXIV.2511.20785. URL https://doi.org/10.48550/arXiv.2511.20785.

Zuhao Yang, Kaichen Zhang, Sudong Wang, Keming Wu, Zhongyu Yang, Bo Li, Xiaojuan Qi, Shijian Lu, Xingxuan Li, and Lidong Bing. Paravt: Taming the tool prior paradox for parallel tool use in agentic video reinforcement learning. *CoRR*, abs/2605.20342, 2026. doi: 10.48550/ARXIV.2605.20342. URL https://doi.org/10.48550/arXiv.2605.20342.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f22a9af8dbb348952b08bd58d4734b50-Abstract-Conference.html.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *CoRR*, abs/2601.23224, 2026. doi: 10.48550/ARXIV.2601.23224. URL https://doi.org/10.48550/arXiv.2601.23224.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pp. 543–553. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-DEMO.49. URL https://doi.org/10.18653/v1/2023.emnlp-demo.49.

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. *CoRR*, abs/2508.04416, 2025a. doi: 10.48550/ARXIV.2508.04416. URL https://doi.org/10.48550/arXiv.2508.04416.

Hongjie Zhang, Lu Dong, Yi Liu, Yifei Huang, Yali Wang, Limin Wang, and Yu Qiao. Lvbench: A benchmark for long-form video understanding with versatile multi-modal question answering. *Int. J. Comput. Vis.*, 133(11):7726–7747, 2025b. doi: 10.1007/S11263-025-02555-8. URL https://doi.org/10.1007/s11263-025-02555-8.

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. *Trans. Mach. Learn. Res.*, 2025, 2025c. URL https://openreview.net/forum?id=30RAWQVGlx.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025, 2025d. URL https://openreview.net/forum?id=EElFGvt39K.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13691–13701. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01278. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_MLVU_Benchmarking_Multi-task_Long_Video_Understanding_CVPR_2025_paper.html.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. Videobrain: Learning adaptive frame sampling for long video understanding. *CoRR*, abs/2602.04094, 2026. doi: 10.48550/ARXIV.2602.04094. URL https://doi.org/10.48550/arXiv.2602.04094.

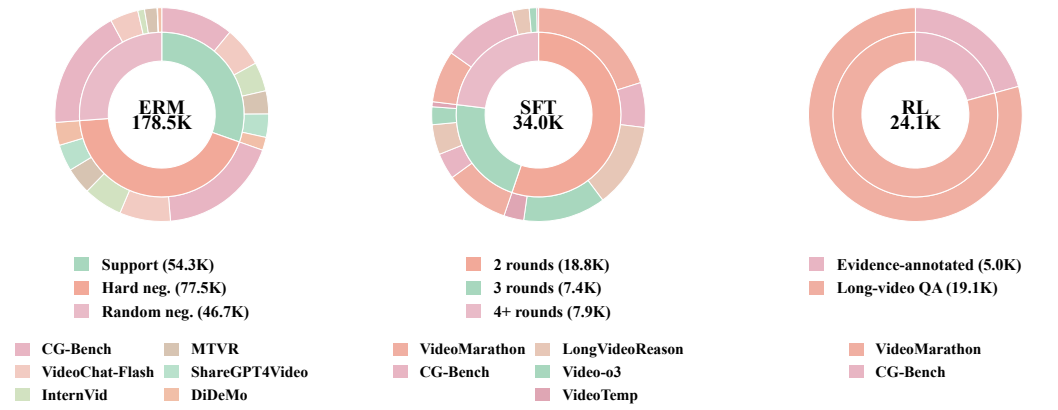
A TRAINING DATA

A.1 ERM TRAINING DATA

The ERM data are organized around the distinction the evidence scorer must learn over accumulated trajectories: whether several retrieved windows jointly contain enough visual information to answer the question. We collect temporally annotated questions from CG-Bench, MultiTaskVideoReasoning (MTVR), VideoChat-Flash, ShareGPT4Video, and the InternVid and DiDeMo subsets of VideoMind (Chen et al., 2025a; Zhang et al., 2025a; Li et al., 2025b; Chen et al., 2024; Liu et al., 2025). For each question, evidence-overlapping windows form the support pool. Hard negatives preserve locally relevant context while excluding the annotated event, and random negatives are drawn from elsewhere in the video.

ERM training instances contain sampled video windows without an initial global overview. These windows are combined into matched contexts. Positive contexts combine the supporting windows with hard or random distractors. Negative contexts preserve the same context size and distractor composition but omit or replace the supporting content. When annotations identify multiple evidence intervals, partial contexts containing only a subset of those intervals remain negative until the complete supporting set is present. We additionally construct nested context pairs in which the final window either completes the required evidence or contributes only another distractor. Varying the number and order of constituent windows prevents shortcuts based on context length or evidence position. This construction exposes the ERM to the same central question encountered during policy optimization: whether the evidence accumulated so far is jointly sufficient and whether the latest observation changes that state.

We merge sources with different video domains and evidence granularity so that the ERM does not specialize to a single annotation style. Splits are constructed at the video level, with held-out and downstream evaluation videos excluded before merging. Corrupted or over-budget visual inputs are removed during the final integrity check. Figure 6a reports the resulting composition.



(a) ERM Training Data **(b)** VEPO SFT Data **(c)** VEPO RL Data

Figure 6: Training data distributions. Inner rings show the constituent window types used to assemble multi-window ERM contexts in (a), the number of search rounds in (b), and the RL prompt type in (c). Outer rings decompose each category by data source.

A.2 VEPO TRAINING DATA

SFT Data. Cold-start SFT is used to teach the interaction grammar and the search–observe–answer workflow. Strong proprietary teachers generate multi-round trajectories over questions from VideoMarathon, CG-Bench, LongVideoReason, Video-o3, and VideoTemp (Lin et al., 2025; Chen et al., 2025a;b; Zeng et al., 2026; Liu et al., 2026). We retain only trajectories with a correct final answer and multiple search rounds, ensuring that each example demonstrates evidence acquisition instead of immediate termination. Tool requests must align with recorded observations, the terminal

answer must follow the required format, and trajectories containing fabricated observation markers are removed before question-level deduplication. Retaining trajectories of different lengths exposes the model to both continued search under uncertainty and termination after sufficient evidence is acquired.

RL Data. The RL mixture serves a complementary role. It retains the video, question, answer, and a coarse uniformly sampled overview, but removes the teacher interaction trajectory. We combine broad long-video questions from VideoMarathon with evidence-annotated questions from CG-Bench, then let the current policy generate every temporal search and final answer online. Consequently, SFT supplies a valid interaction prior, whereas RL exposes the policy to its own search states and lets VEPO determine which trajectories and search actions merit credit. Figures 6b and 6c summarize the two mixtures.

B Training Details

B.1 Experimental Setup

To evaluate whether aligning optimization with evidence sufficiency improves long-video reasoning, we assess our framework on VideoMME (Fu et al., 2025), MLVU (Zhou et al., 2025), LongVideoBench (Wu et al., 2024), and LVBench (Zhang et al., 2025b), which require policies to actively retrieve and integrate information across minute- to hour-long videos. To isolate the effect of the learning objective, we establish a controlled comparison using a shared Qwen2.5-VL-7B policy (Bai et al., 2025) initialized from a supervised fine-tuning (SFT) model that provides the basic search-and-answer grammar. From this shared starting point, we contrast three distinct optimization signals: direct GRPO isolates the effect of outcome-only supervision by rewarding final-answer correctness alone; supplementing this objective with a generic tool-use reward tests whether heuristic execution incentives can sustain active search; and VEPO applies our proposed trajectory-level semantic sufficiency and action-local evidence credit. We additionally situate these controlled variations against proprietary models, dense open-source video MLLMs, and existing agentic frameworks to establish their relative performance.

B.2 Training Configuration

Table 4 summarizes the optimization and visual-input configurations for all three training stages. The ERM is trained independently and remains frozen thereafter. SFT initializes the Qwen2.5-VL-7B policy with the multi-turn interaction protocol, and VEPO continues from this checkpoint for one epoch of grouped on-policy optimization. During each rollout, the policy receives eight frames from every search call, while each returned window contributes four frames to the accumulated visual context jointly assessed by the frozen ERM. This subsampling accommodates the ERM's visual context limit while retaining observations from all retrieved windows. The ERM score is conditioned on this sampled representation, which may omit details visible in the policy's denser input. The initial overview is excluded from evidence-judge inputs. We use an evidence threshold of $\tau = 0.5$. The trajectory rewards $r_{\text{wrong}} = 0.0$, $r_{\text{corr}} = 0.4$, and $r_{\text{evi}} = 1.0$ impose the ordering wrong answer $<$ unsupported correct answer $<$ evidence-supported correct answer, making visual support determine the preferred form of correctness.

C Supplementary Results

C.1 Dynamic Search Depth

VEPO uses a fixed maximum interaction horizon but does not prescribe a fixed search depth. After every search, the policy decides whether to answer or inspect another temporal region based on the accumulated interaction state. We probe this dynamic stopping behavior by varying the number of frames returned per search turn while keeping the policy and maximum horizon fixed. As shown in Figure 7, accuracy improves and then saturates as each search exposes more visual content, while the realized number of searches steadily decreases. The interaction limit therefore sets the maximum search capacity: VEPO allocates additional turns only when further inspection remains useful.

Table 4: Training configurations for the ERM, SFT, and VEPO.

Configuration	ERM	SFT	VEPO
Backbone	Qwen3-VL-4B	Qwen2.5-VL-7B	Qwen2.5-VL-7B
Learning rate	1×10^{-6}	1×10^{-5}	1×10^{-6}
Global batch size	8	32	256
Epochs	1	1	1
Rollout group size	8	–	8
Max. interaction turns	–	–	15
Frames per RM window	4	–	4
Frames per search call	–	–	8
RM image max. edge	640	–	640

The paired 16-to-32-frame comparison shows that the number of calls adapts to each trajectory. More than half of the questions terminate with fewer searches when each turn returns 32 frames, while answer accuracy remains unchanged. The additional visual content thus substitutes for future interaction on a question-dependent basis. Narrow observations leave unresolved uncertainty and induce deeper search, whereas richer observations allow the policy to stop earlier without sacrificing correctness. This dynamic allocation is the practical advantage of adaptive search: difficult questions can consume additional interaction, while questions resolved early avoid unnecessary calls. The result demonstrates adaptive search depth. Richer observations can still contain more frames despite requiring fewer turns, so reduced search depth does not imply lower total visual computation.

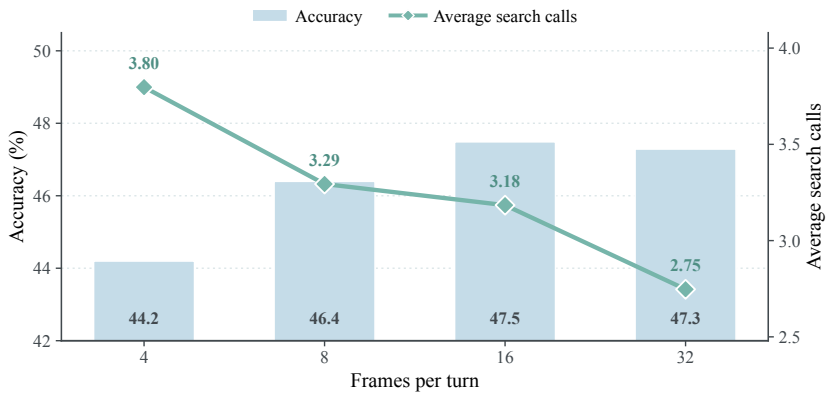


Figure 7: Dynamic search depth on LVBench. Bars report answer accuracy and the curve reports average realized search calls as the per-turn frame budget increases.

C.2 Inference Efficiency

Long-video inference exposes a fundamental efficiency trade-off. Dense inference spends computation on a large uniformly sampled visual context, whereas agentic methods exchange part of this context cost for sequential search. Table 5 compares the resulting accuracy–latency frontier on LVBench. VEPO occupies the high-accuracy end of the Pareto frontier. Relative to dense Qwen2.5-VL-7B inference, it reduces latency by 57% while improving accuracy by 12.9 points, indicating that the smaller visual context offsets the cost of sequential interaction. Video-MTR provides the lowest latency but trails VEPO by 7.1 accuracy points. The remaining methods also fall below VEPO in accuracy, and LongVT and Conan incur even greater latency. The advantage therefore comes from allocating computation to selected temporal regions, reducing the cost of processing a dense visual context while retaining the cost of sequential interaction.

C.3 Generalization across Policy Backbones

To test whether the effect of VEPO transfers beyond the policy used in the main experiments, we apply the same evidence-aware objective and RL data to the Qwen3-VL-4B base model, changing both the model family and parameter scale. Figure 8 reveals a consistent evidence-aligned optimization signature across the two backbones: answer accuracy and supported correctness based on the joint accumulated evidence improve together, while temporal evidence recall and active search remain sustained as task performance rises. Recovering this coupled behavior with a smaller, distinct policy shows that the benefit of evidence credit is not specific to the Qwen2.5-VL-7B initialization.

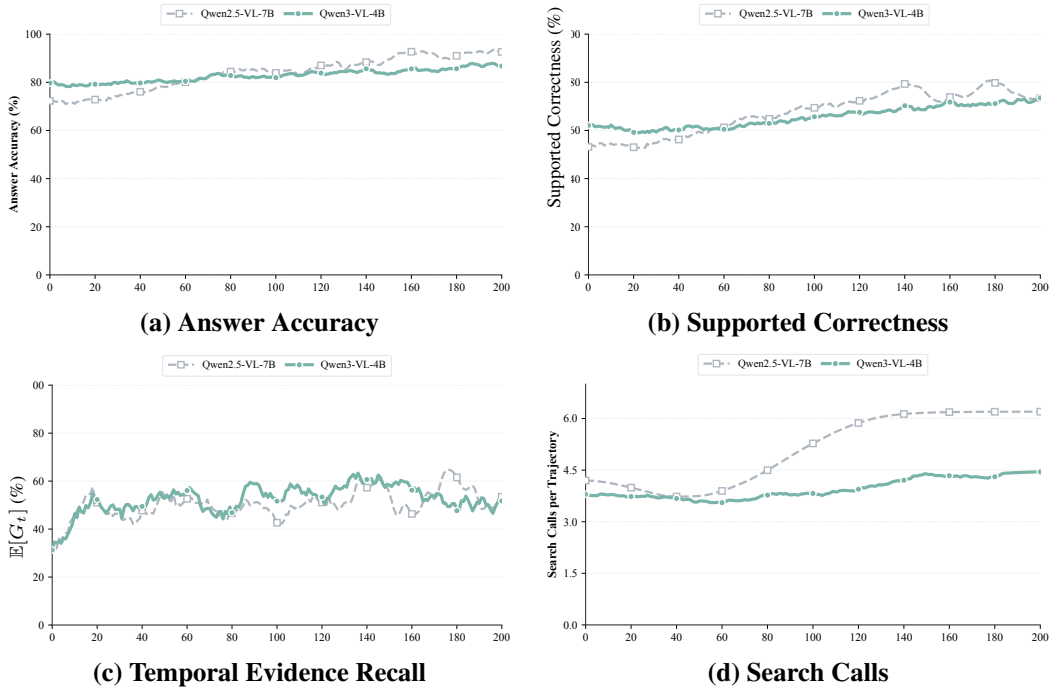


Figure 8: VEPO training dynamics across policy backbones. Supported correctness uses the joint ERM judgment over each trajectory's accumulated observations.

Table 5: Inference efficiency.

Method	Acc. (%)	Latency (s)
Qwen2.5-VL-7B (direct)	34.6	64.9
LongVT	29.4	31.2
Conan	39.7	39.2
Video-MTR	40.4	12.4
VideoTemp-o3	36.8	20.0
TimeSearch-R	38.4	20.5
VEPO	**47.5**	**27.9**

Table 6: Policy-backbone generalization.

Method	MLVU	VideoMME	LongVideo Bench	LVBench
Qwen3-VL-8B	58.3	59.9	52.2	33.1
ParaVT-8B	65.0	62.1	60.4	39.8
VideoBrain-8B	46.9	–	53.3	41.3
Qwen3-VL-4B	65.9	62.9	58.7	44.2
VEPO	**70.7**	**66.3**	**63.5**	**51.1**

Table 6 confirms that this transfer extends to final task performance. VEPO improves the Qwen3-VL-4B base model across all four benchmarks, with the largest gain on LVBench, and the resulting 4B policy surpasses the Qwen3-VL-8B, ParaVT-8B, and VideoBrain-8B results reported in their original papers (Team, 2025b; Yang et al., 2026; Zou et al., 2026). Across both policy families, evidence-credit optimization remains effective under reduced policy capacity, preserving evidence seeking while strengthening long-video reasoning.

Evidence assessment protocol. The indicator G_s records whether the observations acquired through temporal search jointly provide sufficient information to answer the question. The main diagnostics in Figures 2 and 4 and Table 2 use the frozen ERM, with its sufficiency score thresholded at $\tau = 0.5$. The cross-judge evaluation below uses Gemini 2.5 Flash as an alternative model

judge and implements the same semantic sufficiency assessment through majority-vote human judgments. The definition of G_s remains unchanged across these implementations.

Every evaluated policy receives an initial global overview selected by the evaluation protocol. Evidence judges exclude this overview and assess only observations returned by the policy's temporal searches. The ERM prefix scores likewise use only search-acquired observations. Supported correctness therefore measures whether the policy's retrieved evidence supports its answer. The overview can still help the policy answer correctly, so an unsupported prediction under this protocol may draw on visual information outside the judge's input.

C.4 Cross-Judge Validation of Evidence Support

To test whether the comparison depends on the training-time evidence judge, we independently reassess evidence support on a sample of 500 trajectories from LVBench using Gemini 2.5 Flash (Team, 2025a) and labels collected from five blinded human experts. Each judge receives the question and the complete set of visual observations retrieved along the trajectory, presented jointly without the method identity, ERM score, agent reasoning, or predicted answer. Table 7 reports the same supported-correctness criterion, $C = G_s = 1$, under ERM, Gemini, and human-consensus judgments of the accumulated multi-window evidence. Against majority-vote human labels, Gemini attains 93.4% support-detection accuracy, Cohen's $\kappa = 0.768$, and F1$= 0.809$. These results support its use as an independent comparative diagnostic, with human judgments providing a reference for its remaining errors.

Table 7: Cross-judge evaluation of joint multi-window evidence support on LVBench.

Method	**Acc.**	**Supported Correctness ($C = G_s = 1$)**		
		ERM	**Gemini 2.5 Flash**	**Human Experts**
LongVT	29.4	11.8	10.5	11.2
Conan	39.7	0.0	0.0	0.0
Video-MTR	40.4	14.1	12.8	13.5
VideoTemp-o3	36.8	13.6	11.8	12.1
TimeSearch-R	38.4	17.3	11.5	12.0
VEPO (Ours)	**47.5**	**36.3**	**31.8**	**33.2**

D Auxiliary ERM Evaluation

D.1 Evidence Recognition by the ERM

VEPO relies on ERM scores being sensitive to answer-supporting visual content. Table 8 examines this ability outside policy optimization. On unseen videos, we compare the original Qwen3-VL-4B model with the trained ERM in distinguishing supporting windows from irrelevant visual context. We then use each model to rank candidate windows under the same frame budget and pass the selected frames to a separate Qwen2.5-VL-7B-Instruct answering model. This design tests whether the learned evidence signal identifies visual content that remains useful beyond the ERM's own classification task. Evaluation details are provided below, while Appendix C.4 independently compares joint multi-window judgments with Gemini and blinded human labels.

The trained ERM improves every evidence-recognition metric, and the windows it selects raise answer accuracy across all four benchmarks under the same frame budget, with the largest downstream gain on LVBench. The improvement is moderate but consistent, showing that ERM training produces an evidence signal that transfers from the recognition task to a separate answering model. In VEPO, this signal is applied jointly to the accumulated observations that define the grounded trajectory reward and local evidence credit.

Table 8: Auxiliary single-window evidence recognition and downstream video understanding using frames selected by the base model and trained ERM.

(a) Evidence Recognition

Model	Acc.	F1	AUROC
Qwen3-VL-4B	73.2	71.8	82.0
ERM (ours)	**80.6**	**80.3**	**87.5**

(b) Downstream Video Understanding

Model	MLVU	VideoMME	LongVideo Bench	LVBench
Qwen3-VL-4B	64.6	59.8	54.8	47.3
ERM (ours)	**66.1**	**61.0**	**56.5**	**50.6**

D.2 Evaluation Protocol

The policy-time ERM jointly evaluates all observations accumulated along a trajectory. The following experiments isolate its single-window recognition and ranking capability as an auxiliary diagnostic. We use the CG-Bench validation partition, which has no question or video overlap with ERM training. Each question contributes the positive window with maximum overlap and the farthest sampled random negative window, producing a balanced binary task. Each evaluation window is represented by four sampled frames. Calibration and test examples are split at the video level.

The held-out test set contains 500 windows and is constructed without test predictions or correctness. The sampling configuration is selected using annotations and calibration data only. We retain 125 balanced pairs for threshold calibration and 250 balanced pairs from a video-disjoint test split. For this window-level diagnostic, each candidate is evaluated as the single-observation case $p_\phi(y = 1 \mid q, \{o\})$. A decision threshold is selected for each model by maximizing positive-class F1 on the calibration pairs and is then fixed for test.

For downstream validation, we partition each video into nonoverlapping candidate windows of 16 seconds. The base model and trained ERM score the two-frame observation from each candidate using the same single-observation case. Frames from the windows with highest scores are ordered chronologically and capped at 64 before being passed to Qwen2.5-VL-7B-Instruct. Both selectors are evaluated on the same questions from MLVU, VideoMME, LongVideoBench, and LVBench.

E Qualitative Case Studies

Iterative temporal search. Figure 9 shows that invoking a tool is not itself sufficient: a single broad crop produces an incorrect answer, whereas successive temporal refinement isolates the answer-relevant moment and yields the correct response.

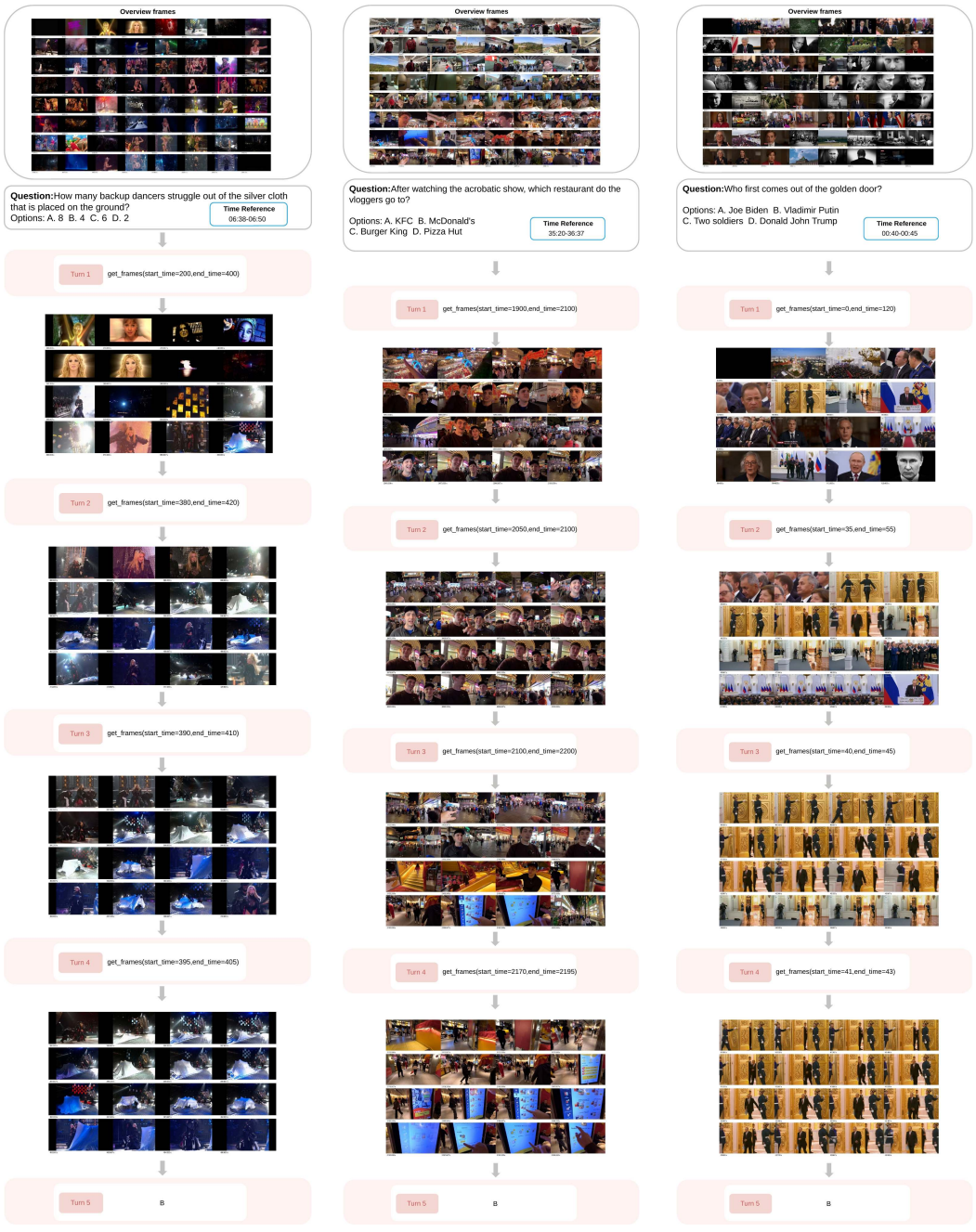
VEPO trajectories. We present five LVBench trajectories produced by the Qwen2.5-VL-7B policy after VEPO training. Following the case taxonomy used in prior agentic long-video analysis, the examples separate three successful search patterns from two diagnostically distinct failures. Each trajectory begins with the same 64-frame global overview; the labels beneath these frames are the actual timestamps supplied in the model prompt. Every subsequent panel shows the interval requested by one `get_frames` action and the timestamped frames returned to the policy. The blue box gives the human-annotated evidence interval and is included to aid interpretation and is not provided to the policy.

Successful evidence seeking. **S1 (Counting)** requires resolving how many dancers emerge from the cloth. The policy first covers a broad concert interval and then successively refines from 380–420 seconds to 395–405 seconds, where the four dancers can be counted; it answers B correctly as illustrated in the left panel of Figure 10. **S2 (Post-condition)** asks what happens after an acrobatic show. The policy follows the vlog beyond the show, re-centers its searches on the later mall sequence, and identifies McDonald's, again answering B correctly as illustrated in the center panel of Figure 10. **S3 (Recovery)** illustrates recovery from an initially broad window: after inspecting 0–120 seconds, the policy narrows to 35–55, 40–45, and finally 41–43 seconds, resolving that Vladimir Putin exits the golden door first and selecting B as illustrated in the right panel of Figure 10.



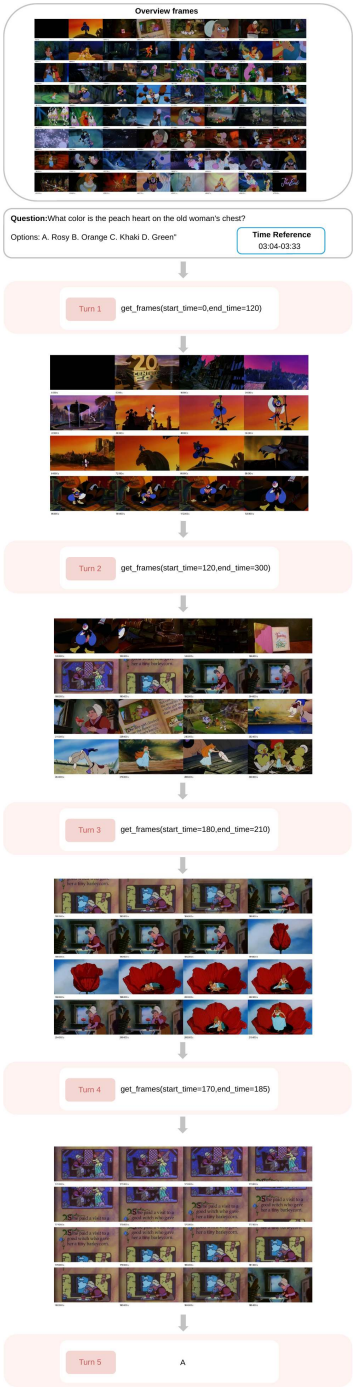
Figure 9: A representative contrast between superficial and iterative tool use. A single broad crop leads to an incorrect answer, whereas successive temporal refinement isolates the answer-relevant moment and yields the correct answer.

Failure modes. **F1 (Fine-grained perceptual error)** retrieves the relevant old-woman and chest sequence, including repeated close views around the annotated interval, but misreads the peach-heart color and predicts A instead of the ground-truth C as illustrated in the left panel of Figure 11. The policy retrieves the relevant event but fails to interpret the acquired evidence correctly. In contrast, **F2 (Off-target retrieval)** never reaches the annotated event at 53:26–53:52: all four searches remain within the first three minutes of the wildlife video. The policy consequently answers from unrelated lion and deer shots, predicting C instead of the ground-truth D as illustrated in the right panel of Figure 11. Together, these cases distinguish whether an error originates in visual interpretation after a useful retrieval or in the temporal search trajectory itself.

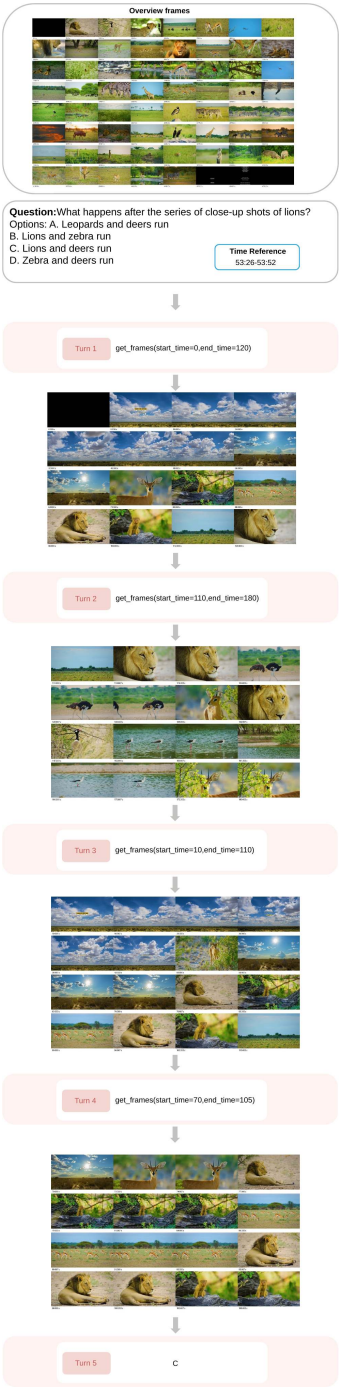


S1: Counting. **S2: Post-condition.** **S3: Recovery.**

Figure 10: Successful VEPO trajectories on LVBench. S1 resolves counting through temporal refinement, S2 follows the video beyond a preceding event, and S3 recovers by repeatedly narrowing the search interval.



F1: Fine-grained perceptual error.



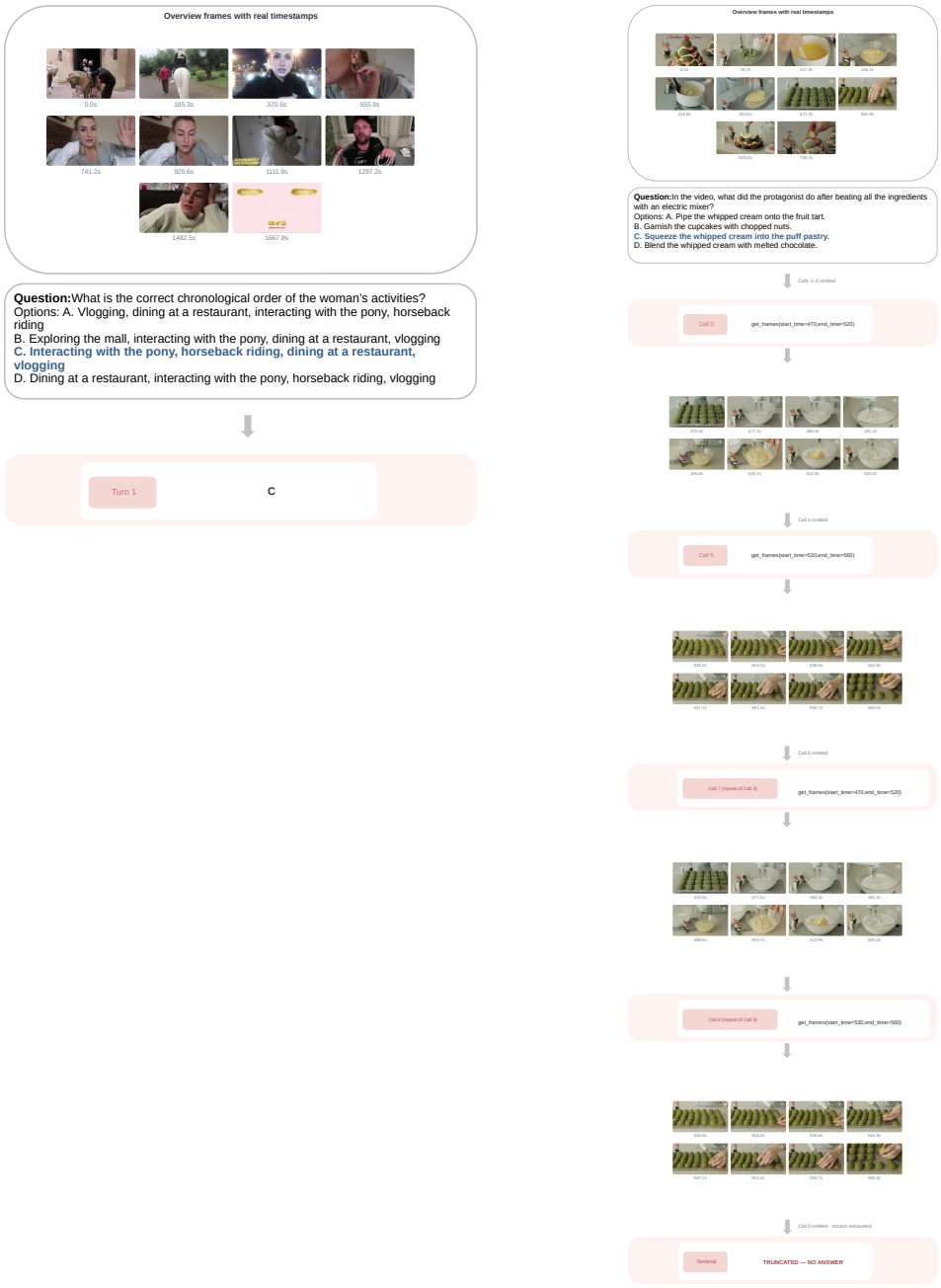
F2: Off-target retrieval.

Figure 11: Two distinct failure modes. F1 reaches the relevant evidence but misclassifies a fine-grained color attribute, whereas F2 never visits the annotated event.

E.1 TRAINING-TIME REWARD SHORTCUTS

Individual training trajectories illustrate how policies can obtain reward without completing evidence acquisition or the answering task, consistent with the training dynamics in Section 2. Under outcome-only GRPO, a policy can answer directly from the initial global overview and receive the full outcome reward when correct, despite executing no temporal search actions as illustrated in the left panel of Figure 12. The objective does not check whether the overview provides sufficient visual support, giving the same reward to this trajectory and a correct answer supported by targeted retrieval. This permits correct predictions to be reinforced without reinforcing additional evidence acquisition, consistent with rising answer accuracy and declining search activity.

Supplementing this objective with a per-call tool reward permits an opposing shortcut. A policy can obtain positive return by repeatedly searching, including revisiting identical temporal windows, even when it exhausts the interaction horizon without producing a final answer as illustrated in the right panel of Figure 12. The corresponding training run shows increased interaction volume and truncation alongside declining answer accuracy. Together, these cases illustrate that outcome-only rewards permit success without checking evidence sufficiency, while tool-use incentives permit rewarded interaction without task completion. Evaluating whether accumulated observations jointly support a correct answer would distinguish successful evidence acquisition from these rewarded behaviors.



Outcome-only Direct GRPO: zero-call termination. **Per-call Tool Reward: repeated search and truncation.**

Figure 12: Training-time reward-hacking trajectories. Left: outcome-only Direct GRPO receives full return after answering from the overview without calling the tool. Right: a per-call tool bonus rewards nine valid searches, including two exact repeated windows, even though the rollout truncates without an answer. Middle calls are explicitly omitted for readability.

F PROMPTS

Following the presentation used in the reference appendix, we reproduce the textual interfaces used by VEPO below. The policy system prompt is shared by cold-start SFT, RL, and benchmark inference. Braced fields denote values filled for each example, and each <image> token is paired with the frame at the preceding timestamp. We separately report the frozen ERM prompt used for training-time evidence credit and the blinded cross-judge prompt used only for the supplementary Gemini evaluation. Optional force-answer fallbacks were disabled in the formal evaluation and are therefore not included.

F.1 SYSTEM PROMPT FOR THE VIDEOQA POLICY

VideoQA Tool-Agent System Prompt

```
You are a video question-answering assistant. To look at any part of
the video, call get_frames:
<tool_call>{"name": "get_frames", "arguments": {"start_time":
<seconds>, "end_time": <seconds>}}</tool_call>
It returns frames from that window, each prefixed with its timestamp <T
seconds>. Think inside <think> </think>, then give the final answer
inside <answer> </answer>.
```

F.2 POLICY INPUT AND TOOL-OBSERVATION TEMPLATES

The initial user message combines a uniform global overview with the task. The same serialization is used in SFT and RL data; formal evaluation increases the number and resolution of overview frames without changing the text template.

Initial User Prompt Template

```
The video is {duration_seconds} seconds long. Overview frames across
the whole video:
<{overview_time_1} seconds><image> ... <{overview_time_N}
seconds><image>
{question}
Options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
```

After a valid get_frames call, the environment returns the sampled frames in chronological order using the following user-message template.

Tool Observation Template

```
Here are the sampled frames:
<{frame_time_1} seconds><image> ... <{frame_time_M} seconds><image>
```

If a policy turn contains neither a parseable tool call nor a final answer, the environment returns the following correction message and allows interaction to continue.

Invalid-Action Correction Message

```
No valid tool call. Use <tool_call>{"name":"get_frames","arguments":{"␣
start_time":S,"end_time":E}}</tool_call> to view a window, or give
<answer>X</answer> to finish.
```

F.3 PROMPT FOR THE EVIDENCE REWARD MODEL

The frozen Qwen3-VL-4B ERM receives, in a single user turn, the complete set of video windows retrieved through search actions up to the current interaction step, excluding the initial overview.

Frames are grouped by retrieval window and ordered chronologically within each group, so the model judges whether their combined visual content is sufficient. It produces one token, and evidence confidence is computed from the aggregated Yes/No token probabilities.

Multi-Window Evidence-Sufficiency Prompt

```
You are a strict visual-evidence sufficiency judge for long-video
question answering.
Given a question and frames sampled from retrieved video windows,
decide whether their combined visual content provides sufficient
evidence to determine the correct answer.
Judge all retrieved video windows jointly. Return No when the relevant
event is absent, incomplete, unreadable, or ambiguous, or when the
windows show only similar objects, repeated title cards, background
scenes, or related context outside the evidence moment.
Answer with exactly one token: Yes or No.
Do not output punctuation, explanation, or any other token.

Question:
{question_and_options}

Retrieved video windows:
Window 1: {start_time_1}s to {end_time_1}s: <image> ... <image>
...
Window K: {start_time_K}s to {end_time_K}s: <image> ... <image>
Answer:
```

F.4 Prompt for Blinded Cross-Judge Validation

This prompt is used only for the independent Gemini 2.5 Flash validation in Appendix C.4. Method identity, ERM score, agent reasoning, and the model's predicted answer are not exposed to the judge.

Gemini Evidence-Sufficiency Judge Prompt

```
[System]
You are an independent, blinded visual-evidence sufficiency judge.
Your task is not to evaluate an agent or its reasoning. Decide only whether the retrieved
visual observations jointly contain enough visible evidence to determine the correct answer
to the supplied multiple-choice question.

Images are ordered by retrieved window and then chronologically within each window. The
initial overview is not included. Mark supported=true only when the visible observations
establish one answer option with high confidence. Mark supported=false when the relevant
event is absent, incomplete, unreadable, ambiguous, or would require narrative priors,
outside knowledge, or guessing. Do not assume that nearby frames contain an event that is
not visible.

Return one JSON object and no markdown:
{"supported": true or false, "answer_choice": "A" or "B" or "C" or "D" or null,
"rationale": "one short evidence-based sentence"}

[User]
Question:
{question_and_options}
Judge whether the retrieved observations below are jointly sufficient to determine the
correct option.
Retrieved window {window_index}: {start_time}s-{end_time}s
Window {window_index}, frame {frame_index}: <image> ...
```