WORTH: DECISION-ALIGNED CREDIT ASSIGNMENT FOR EVIDENCE ACQUISITION IN LONG VIDEO AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic video understanding has emerged as a promising paradigm for long-video reasoning, wherein an agent first perceives a sparsely sampled overview and then selectively retrieves evidence from task-relevant temporal segments. However, existing agents are typically optimized via outcome-based reinforcement learning with a single trajectory-level reward, which neither evaluates whether a given acquisition justifies its cost nor attributes the final outcome to individual observations. We term this deficiency the *acquisition–attribution mismatch*, and our systematic analysis shows that it drives agents to acquire evidence indiscriminately while most of the acquired observations contribute nothing to the final answer. To address this, we propose **WORTH** (**W**hether **O**bservation **R**etrieval **T**ruly **H**elps), a decision-aligned reinforcement learning framework that decouples acquisition from attribution via two dedicated credit signals. An *acquisition credit* estimates the ex-ante value of retrieval as the cost-adjusted gain of acquiring over answering immediately, computed through paired continuations from the same state. An *observation credit* estimates the ex-post contribution of each retrieved observation as its exact Shapley value, obtained by re-answering from every subset of acquired observations. Both credits are derived from the policy itself and replace the trajectory-level advantage exclusively on the tokens corresponding to their respective decisions. Experiments demonstrate that WORTH achieves state-of-the-art performance among most video agents across multiple long-video understanding and temporal grounding benchmarks, improving accuracy by 2.2% over outcome-based RL baselines while consuming 67% fewer tokens. Code, data, and models will be released.

1 INTRODUCTION

Long video understanding remains a fundamental challenge for multimodal large language models (MLLMs) (Bai et al., 2025a; An et al., 2026), since the evidence required by a question is typically confined to a few brief moments within hours of footage. This sparsity leaves two unsatisfactory options. Processing the entire video at high resolution spends most of the computation on irrelevant content, whereas sparse uniform sampling inevitably misses the brief moments that matter. To resolve this dilemma, agentic video understanding has emerged as a promising paradigm (Yang et al., 2026a; Jain et al., 2026; Zeng et al., 2026; Liu et al., 2026c), in which a model first perceives a sparsely sampled overview, then acquires additional evidence from selected temporal segments, and finally answers based on the collected evidence. The central premise of this paradigm is that evidence should be acquired selectively, namely, only when the expected benefit justifies the additional cost.

However, existing agents are predominantly trained by outcome-based reinforcement learning (Shao et al., 2024; Yang et al., 2026a;b), where a single trajectory-level reward is uniformly assigned to all decisions in a trajectory, including whether to acquire evidence, what to acquire, and how to answer. Consequently, this reward can neither assess whether an acquisition justifies its cost nor attribute the outcome to individual observations. For instance, a trajectory that reaches the correct answer after unnecessary acquisitions receives the same reward as a trajectory that answers directly from the overview. To examine whether this deficiency affects existing agents in practice, we conduct a systematic analysis of representative agents (Figure 1), and confirm that the deficiency is pervasive. These findings raise three fundamental questions: (1) ***Why do existing agents acquire evidence***

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.

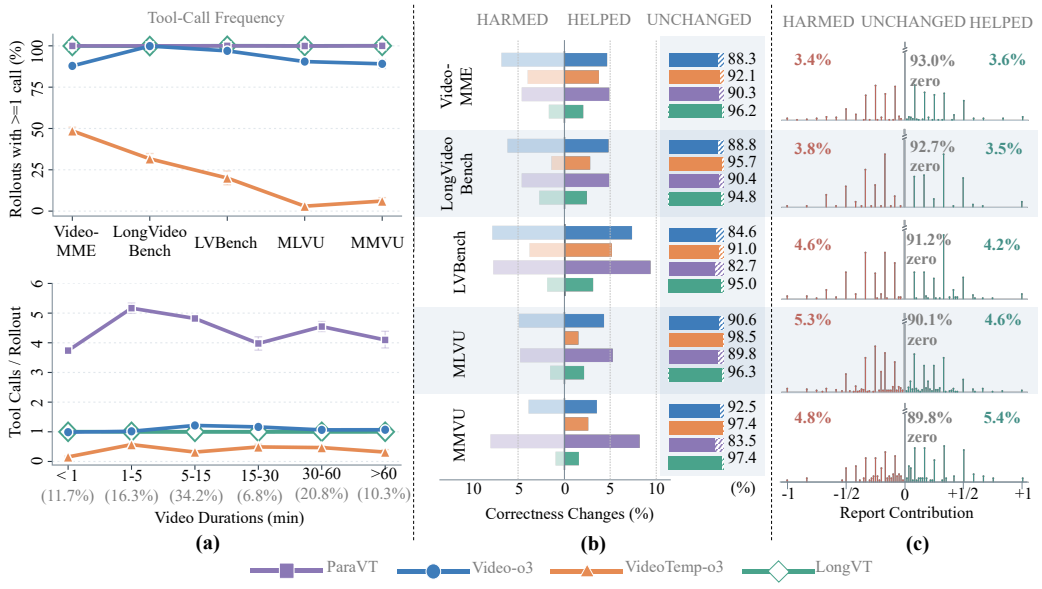


Figure 1: **Systematic analysis of representative agents.** **(a)** Proportion of rollouts with at least one tool call per benchmark (top), and calls per rollout by video durations (bottom). **(b)** Proportion of questions whose correctness is harmed (light) or helped (dark) by the first acquisition, with the unchanged proportion listed alongside. **(c)** Distribution of the exact Shapley value of each observation acquired in parallel by ParaVT (Yang et al., 2026b).

on nearly every question, regardless of the video duration (Figure 1a)? (2) Why does the first acquisition change only 2–17% of the answers, and harm them almost as often as it helps (Figure 1b)? (3) Why do 92% of the observations acquired in parallel contribute nothing to the final answer (Figure 1c)?

Through an in-depth analysis of these questions, we uncover that all three phenomena originate from a structural deficiency of trajectory-level supervision, rather than from the capability of individual agents. In particular, deciding whether to acquire evidence and identifying which observations contribute are two distinct credit assignment problems, which arise at different stages of a trajectory. The former concerns the ex-ante value of acquisition, which is determined before any evidence is acquired. In contrast, the latter concerns the ex-post contribution of each observation, which can only be determined after the acquisition is executed. A single trajectory-level reward inevitably conflates these two quantities, and we term this problem the *acquisition–attribution mismatch*. Consequently, an acquisition is reinforced whenever the trajectory succeeds, even when the answer is already correct without it, while every observation receives the same credit irrespective of its actual contribution.

Motivated by these insights, we propose **WORTH** (**W**hether **O**bservation **R**etrieval **T**ruly **H**elps), a decision-aligned reinforcement learning framework that decouples acquisition from attribution through two dedicated credits. To make both credits measurable, WORTH first adopts a direct-action interface, in which the policy either answers or requests temporal segments at each turn, and a reader inspects each requested segment and returns a textual observation. Since answering is an explicit action under this interface, the outcome of answering immediately from any state can be obtained from the policy itself without an auxiliary model. WORTH then supervises each decision with its dedicated credit. The acquisition credit estimates the ex-ante value of acquisition through paired continuations that either acquire evidence or answer immediately from the same state, with the evidence cost deducted. The observation credit, in turn, estimates the ex-post contribution of each observation as its exact Shapley value, which is computed by re-answering from every subset of the acquired observations. Finally, each credit replaces the trajectory-level advantage only on the tokens of its corresponding decision. In this sense, the supervision of WORTH is aligned with each decision rather than with the entire trajectory, which we refer to as decision-aligned.

Our contributions are summarized as follows. **(i)** We conduct a systematic analysis of existing video agents and identify the *acquisition–attribution mismatch*, a structural deficiency of trajectory-level supervision that results in habitual evidence acquisition and leaves the contribution of individual

observations unsupervised. **(ii)** We propose WORTH, a decision-aligned reinforcement learning framework, which estimates the ex-ante value of acquisition and the ex-post contribution of each observation with two dedicated credits and assigns each credit only to the tokens of its corresponding decision. **(iii)** Extensive experiments on multiple benchmarks demonstrate that WORTH achieves SOTA performance among most video agents and outperforms outcome-based reinforcement learning at a substantially lower evidence cost.

2 Diagnosing the Acquisition–Attribution Mismatch

This section analyzes the three questions raised in Section 1, with further details in Appendix B.

2.1 Formalization and Diagnostic Setup

Setup. We analyze four representative video agents, namely Video-o3 (Zeng et al., 2026), VideoTemp-o3 (Liu et al., 2026c), LongVT (Yang et al., 2026a), and ParaVT (Yang et al., 2026b), with their released weights, prompts, and tools (Appendix B.1).

Notation. Let h denote the state that consists of the question and the overview. Given h, an agent π either answers immediately or acquires evidence. In the latter case, it issues a plan P of requests for temporal segments and then answers based on the returned observations O. To isolate the effect of an acquisition, we take the state h right before the first acquisition of each agent and let the agent answer under three conditions. Formally, $a_A \sim \pi(\cdot \mid h)$ is the answer without acquisition, $a_B \sim \pi(\cdot \mid h, P, \varnothing)$ is the answer after the requests are issued but with the observations withheld, and $a_C \sim \pi(\cdot \mid h, P, O)$ is the answer with the observations. Let $c_A, c_B, c_C \in \{0, 1\}$ denote the correctness of the three answers. Accordingly, the change in correctness caused by an acquisition decomposes exactly into a request term and an evidence term:

$$\delta_{\text{acq}} \;=\; c_C - c_A \;=\; \underbrace{(c_B - c_A)}_{\delta_{\text{req}}} + \underbrace{(c_C - c_B)}_{\delta_{\text{evi}}}. \tag{1}$$

Diagnostics. Since a trajectory-level reward observes only c_C, we introduce three diagnostics to quantify what this reward overlooks. The unearned acquisition rate U_{acq} is the proportion of successful acquisitions whose answers are already correct before the acquisition. The request share ρ_{req} is the proportion of answer changes that already occur before any evidence is observed:

$$U_{\text{acq}} = \Pr\big(c_A = 1 \mid c_C = 1\big), \qquad \rho_{\text{req}} = \Pr\big(c_B \neq c_A \mid c_C \neq c_A\big). \tag{2}$$

Moreover, for a plan with K observations, we let the agent re-answer from every subset of the observations and compute the exact Shapley value ϕ_i of each observation. The unearned credit rate $U_{\text{obs}} = \Pr(\phi_i \leq 0 \mid c_C = 1)$ is the proportion of observations reinforced by a successful trajectory without contributing to the answer. Since the Shapley values sum to $c_C - c_B$, the outcome observed by the reward decomposes into three levels:

$$c_C \;=\; c_A \;+\; \delta_{\text{req}} \;+\; \textstyle\sum_i \phi_i, \tag{3}$$

whose three terms represent the prior knowledge of the agent, the request term, and the contributions of the individual observations. While a trajectory-level reward depends only on their sum c_C, the three questions in Section 1 address these terms individually. Accordingly, Sections 2.2 to 2.4 analyze each term in turn.

2.2 Why Do Agents Acquire Evidence Indiscriminately?

Phenomenon. Regarding question (1), LongVT and ParaVT acquire evidence on over 99.9% of the questions in every benchmark, and Video-o3 on 88–100%, regardless of the video duration (Figure 1a). Only VideoTemp-o3 acquires evidence selectively. However, most of these acquisitions are unnecessary, as U_{acq} reaches 90–96% for all agents (Table 1a).

Cause. We attribute this behavior to two properties of trajectory-level supervision. First, the reward reinforces an acquisition whenever the trajectory succeeds. Since most successful trajectories do not depend on the acquisition, the decision to acquire is mainly reinforced by the prior knowledge

Table 1: **Diagnostics of existing video agents.** Panel (a) probes the first acquisition of each agent on five benchmarks, and panel (b) partitions the observations and the plans of ParaVT by the sign of the exact Shapley value ϕ_i. All values are in %, and gray marks a net effect whose 95% confidence interval includes zero.

(a) First acquisition of each agent

Agent	Acquisition decision			Effect on the answer			Source of the effect		
	Acq.	U_{acq}	Mixed	Helped	Harmed	δ_{acq}	ρ_{req}	δ_{req}	δ_{evi}
Video-o3	92.2	91.0	43.8	5.1	6.3	−1.24	64	−0.54	−0.70
VideoTemp-o3	25.6	93.1	71.0	3.7	3.3	+0.42	48	+0.42	0.00
LongVT	99.9	95.6	0.6	2.3	1.8	+0.49	37	−0.05	+0.54
ParaVT	100.0	89.9	0.3	6.1	5.6	+0.54	75	+0.70	−0.17

(b) Observation credit of ParaVT

Scope	Shapley value ϕ_i		
	> 0	$= 0$	< 0
All observations	4.1	91.6	4.3
in correct plans	4.9	93.4	1.7
in wrong plans	2.8	88.7	8.5
Entire plans	2.4	94.9	2.7

of the agent, *i.e.*, the term c_A in Eq. (3). Second, the supervision on this decision vanishes as the agent converges, because the expected group-relative advantage on the decision scales with $p \cdot (1-p)$ at an acquisition rate p. At the measured acquisition rates, only 0.6% and 0.3% of the groups of eight rollouts would contain both choices for LongVT and ParaVT (Table 1a). Moreover, multiple agents are trained with auxiliary bonuses for executing tool calls (Table B.1), which further encourage acquisition.

Finding 1. A trajectory-level reward reinforces acquisitions regardless of their necessity, and the signal it provides on the acquisition decision disappears once every rollout acquires.

2.3 Why Does Acquisition Yield Little Net Gain?

Phenomenon. Regarding question (2), the first acquisition alters the correctness of only 2–17% of the answers (a_A vs. a_C, Figure 1b), and the numbers of helped and harmed answers are nearly equal for every agent (Table 1). Consequently, the net effect δ_{acq} lies within $[-1.2\%, +0.5\%]$.

Cause. We attribute this phenomenon to a trajectory-level reward that credits an acquisition as a whole and therefore cannot separate the effect of the requests from the effect of the returned evidence, namely the two terms of Eq. (1). Conditioned only the requests, the intermediate answer a_B naturally isolates the request effect. A large share of the changes already appears in a_B, as ρ_{req} reaches 37–75% for these agents (Table 1a). Since a request depends only on h, it carries no new information about the video. The changes that it induces are therefore random perturbations of uncertain answers, harming as often as they help (δ_{req} is insignificant for all agents). The returned evidence, in contrast, is the only source of new information but exhibits no consistent effect. Its net effect δ_{evi} remains marginal, ranging $[-0.7\%, +0.5\%]$ with inconsistent signs across agents. Furthermore, accessing this evidence fails to improve the correction rate of initially incorrect answers (47–53% vs. 48–51%, Appendix B.4). Because the reward assigns the same credit to both effects, it provides no incentive for the agent to ground its answer in the returned evidence.

Finding 2. A trajectory-level reward credits an acquisition as a whole, so that the answer changes are dominated by random perturbations induced by the requests, whereas the returned evidence exhibits no consistent effect.

2.4 Why Do Most Observations Contribute Nothing?

Phenomenon. Regarding question (3), we compute the exact Shapley value of every observation in the 7,750 parallel plans of ParaVT (Figure 1c, Table 1b). Among all observations, 91.6% have no effect on the answer, and the remaining ones are harmful almost as often as helpful. Even in the plans that end correctly, only 4.9% of the observations contribute to the answer, yielding a U_{obs} of 95%. At the plan level, the observations jointly change the correctness of the answer in only 5.1% of the plans.

Cause. We attribute this phenomenon to the fact that a trajectory-level reward assigns the same credit to all observations of a plan. This credit is misaligned with the contributions, which are neither equal nor independent. The advantage agrees in sign with the Shapley value for only 6.3% of the observations, and 91.6% receive a non-zero advantage despite a zero contribution. The requests

that retrieve useless observations are therefore reinforced whenever the plan succeeds and never suppressed. Nor can the contributions be identified by examining each observation individually. When a plan helps, the gain is shared by two or more observations in 168 of the 189 plans, which is the intended redundancy of parallel acquisition (Yang et al., 2026b). Consequently, leave-one-out attribution assigns zero to every observation in 295 plans whose observations do contribute. A faithful credit should therefore be signed and averaged over all subsets in which an observation appears (Appendix B.5).

Finding 3. A trajectory-level reward assigns the outcome of a plan uniformly to its observations, although their contributions are sparse, signed, and redundant, so that useless observations are persistently reinforced.

2.5 Structural Root Cause and Remedy

The three findings share a single structural cause. A trajectory-level reward observes only the outcome c_C in Eq. (3) and assigns it to all decisions, although the three terms are determined at different stages of a trajectory. The value of an acquisition is determined before any request is issued and should be measured against answering immediately (a_A). The contribution of an observation, in contrast, is determined only after the acquisition and should be measured against the answer under the same requests (a_B). This conflation of the ex-ante value of acquisition with the ex-post contribution of the observations is the *acquisition–attribution mismatch*. Existing remedies address only one side of this mismatch. Necessity-based rewards (Wang et al., 2026a) compare acquiring with not acquiring but assign the result to the entire trajectory, whereas information-gain rewards (Wang et al., 2026b; Xie et al., 2026) compare the answers before and after an observation under an already issued request and thus overlook the request term.

Therefore, a remedy should satisfy three requirements. First, it should estimate the value of an acquisition for the current policy (Appendix B.6) against answering immediately from the same state, deduct the evidence cost that existing rewards ignore, and remain well-defined even when all rollouts acquire evidence (Findings 1 and 2). Second, it should also assign each observation a signed credit that is averaged over the subsets of observations under the same requests (Finding 3). Finally, it should route the two credits to the tokens of their corresponding decisions, so that the supervision is decision-aligned rather than trajectory-level.

3 Methodology

WORTH satisfies the three requirements of Section 2.5 with one idea: each decision is supervised by a counterfactual comparison that the policy computes about itself (Figure 2). A direct-action interface provides the two readouts that these comparisons require, the answer without acquisition and the answer under any subset of observations, as forced continuations of the policy (Section 3.1). Reinforcement learning then estimates an acquisition credit and an observation credit from these readouts and routes each credit to the tokens of its own decision (Section 3.2).

3.1 Direct-Action Interface and Evidence-Verified Cold Start

Direct-Action Interface. The direct-action interface turns the two decisions of Section 2.1 into explicit actions of one policy. Given a video V and a question q, the policy π_θ receives a timestamped overview $\mathcal{O}(V)$ of uniformly sampled frames instead of the full video. At turn t it acts from the state $h_t = (q, \mathcal{O}(V), P_1, O_1, \ldots, P_{t-1}, O_{t-1})$, and a task reward $R \in [0, 1]$ scores the final answer. At each turn, the policy either answers, which ends the episode, or issues a plan P_t of at most five requests, each a natural-language task paired with a temporal window. A reader then inspects each requested window and returns a textual observation. The observations O_t are appended to the state, for at most three plans per episode. No reasoning trace precedes an action, so the first token of a turn is the decision itself. Consequently, both readouts that Section 2.5 calls for are forced continuations of the policy itself. Through forced generation, setting the first token to `<answer>` at state h yields the direct answer without acquisition, while doing so after a plan with a subset of observations withheld

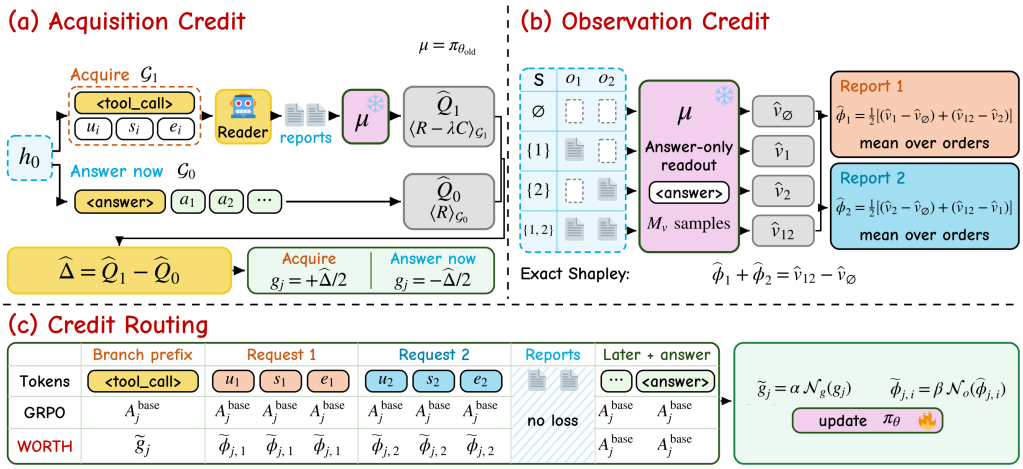


Figure 2: **Overview of WORTH.** Rollouts from the same state either answer immediately or acquire evidence with one or several requests. The acquisition credit contrasts the cost-adjusted returns of the acquiring and the answering rollouts, and the observation credit attributes the gain of a fixed plan to its observations by exact Shapley values over all subsets. Each credit replaces the outcome advantage only on the tokens of its decision before the policy update.

from the context yields the answer conditioned on the remaining subset. Neither readout requires an auxiliary model, a judge, or an annotation.

Evidence-Verified Cold Start. These readouts are informative only if the policy can take both actions at any state and grounds its answers in the returned observations. We therefore initialize the policy by supervised fine-tuning on evidence-verified trajectories. The trajectories are generated from the question and the overview, without access to the answer. A trajectory is accepted only if its final answer is correct, its requested windows cover the annotated evidence, and every claim in its answer is supported by a returned observation (Appendix C.2).

3.2 Decision-Aligned Reinforcement Learning

Intuitively, each credit answers one counterfactual question about a single decision. Had the policy answered immediately instead of acquiring, would the cost-adjusted reward have been higher? Had this observation been withheld, would the answer have changed? Under the direct-action interface both counterfactuals are continuations of the policy, so they can be sampled during training at the price of a few short readouts, and their answers replace the outcome advantage exactly where the corresponding decision is made. For each question, G rollouts are sampled from a policy snapshot μ, scored by R, and given the group-normalized advantage A_j^{base} on every token. WORTH replaces this advantage with an acquisition credit and an observation credit on the tokens of the two decisions and leaves all other tokens unchanged. Both credits are estimated under the same snapshot μ with the reader held fixed, so that each credit compares two behaviors of one policy.

Acquisition Credit. The acquisition credit estimates the ex-ante value of acquisition, namely the expected effect δ_{acq} of Eq. (1) under the current policy, with the evidence cost deducted. Let $C(\tau) \geq 0$ denote the tokens processed by the reader in a continuation τ, and λ convert one unit of cost into task reward. Formally, at a state h the value contrasts the cost-adjusted return of the plan branch with the return of the answer branch:

$$\Delta^{\mu}(h) = \underbrace{\mathbb{E}_{\tau\sim\mu(\cdot|h,\,\text{plan})}\big[R(\tau) - \lambda\, C(\tau)\big]}_{Q_1^{\mu}(h)} - \underbrace{\mathbb{E}_{\tau\sim\mu(\cdot|h,\,\text{answer})}\big[R(\tau)\big]}_{Q_0^{\mu}(h)}. \tag{4}$$

Because both branches start from the same state, the contrast measures the value of acquiring rather than the prior knowledge of the policy. Furthermore, it covers both terms of Eq. (1), since the requests and the returned evidence are both consequences of the same decision. Since all G rollouts of a question start from the same initial state h_0, the group itself samples both branches. Let $\mathcal{G}_1$ and $\mathcal{G}_0$ denote the rollouts whose first action is a plan and an answer. The value is then estimated by the

difference of the branch means,

$$\widehat{\Delta}(h_0) = \frac{1}{|\mathcal{G}_1|} \sum_{j \in \mathcal{G}_1} \left[R_j - \lambda C_j \right] - \frac{1}{|\mathcal{G}_0|} \sum_{j \in \mathcal{G}_0} R_j, \tag{5}$$

in which the request-induced perturbations identified by Finding 2 cancel out on average. When a branch has fewer than a minimum of m members, the missing continuations are generated by forcing the first action and used for scoring only, so that $\widehat{\Delta}$ remains defined even when all rollouts acquire. In contrast, the group-relative advantage of GRPO on this decision scales with $p \cdot (1-p)$ for an acquisition probability p, and vanishes for such one-sided groups (Appendix B.3).

Observation Credit. The observation credit estimates the ex-post contribution of each observation, namely its Shapley value ϕ_i of Section 2.1 under the current policy. Consider a rollout whose first plan P returns the observations $O = (o_1, \ldots, o_K)$, and let $[K] = \{1, \ldots, K\}$. The value of a subset $S \subseteq [K]$ is the expected reward of the answer read out with only the observations in S, *i.e.*, $v(S) = \mathbb{E}_{a \sim \mu(\cdot \mid h, P, O_S)}[R(a)]$, and the contribution of observation i is

$$\widehat{\phi}_i = \sum_{S \subseteq [K] \setminus \{i\}} \frac{|S|! \cdot (K - |S| - 1)!}{K!} \cdot \Big[\widehat{v}(S \cup \{i\}) - \widehat{v}(S) \Big]. \tag{6}$$

Each withheld observation is replaced by a placeholder without re-running the reader, and all $2^K \leq 32$ subsets are enumerated with M_v readouts each, so that the attribution is exact rather than sampled. Because the sum in Eq. (6) ranges over every subset, an observation is credited by its average marginal contribution over all combinations of the other observations. Redundant observations therefore share the gain instead of cancelling, and a misleading observation receives a negative credit, which are the two properties that Finding 3 requires.

Credit Routing. Each credit replaces the outcome advantage only on the tokens that commit to its decision. For rollout j, the decision is committed by its branch prefix $\mathcal{B}_j$, including the first `<tool_call>` or `<answer>` tag. Request i of the first plan occupies the tokens $\mathcal{U}_{j,i}$. The acquisition credit enters with the sign of the branch, $g_j = +\widehat{\Delta}_j/2$ for a rollout that plans and $g_j = -\widehat{\Delta}_j/2$ for one that answers, so that the two branches differ by exactly $\widehat{\Delta}_j$. The token-level advantage is

$$A^{\mathrm{W}}_{j,k} = \begin{cases} \alpha \cdot \mathcal{N}_g(g_j), & k \in \mathcal{B}_j, \\ \beta \cdot \mathcal{N}_o(\widehat{\phi}_{j,i}), & k \in \mathcal{U}_{j,i}, \\ A^{\mathrm{base}}_j, & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{N}_g$ and $\mathcal{N}_o$ scale each credit by its root mean square over the batch. No mean is subtracted, so the sign and the exact zeros of a credit survive, whereas the mean-centering of GRPO would erase both. Observation tokens receive no loss, and all remaining tokens, including the answer, keep the outcome advantage. The policy is updated with the clipped objective of GRPO on the natural rollouts only, and the forced continuations and subset readouts enter training solely through $\widehat{\Delta}$ and $\widehat{\phi}$.

4 Experiments

4.1 Experimental Setup

WORTH is built on Qwen3-VL-8B-Instruct (Bai et al., 2025a). The cold start uses 79k evidence-verified trajectories and 27k reader demonstrations, so that one set of weights serves as the policy and as the reader. RL then runs for one epoch over 5.9k questions with $G = 8$ rollouts per question. Every model receives the same overview of at most 64 frames, decodes greedily, and otherwise runs its released prompts and tools. More details are provided in Appendix C and Appendix D.

4.2 Main Results

Long-Video Understanding. WORTH achieve the best performance on average across long-video understanding under a common overview budget (Table 2), leading on four of the six columns and placing second on Video-MME with subtitles. The most informative comparison is ParaVT, which

Table 2: **Quantitative Results on Five Long Video Understanding Benchmarks.** All the results are reported in percentage. In this table, the default maximum number of frames fed into each solution is 64. The **Best** and Second Best scores are highlighted in **Bold** and Underline.

Models	Param.	VideoMME		LongVideo-Bench	LVBench	MLVU	MMVU
		w/o sub	w/ sub				
Proprietary Models							
GPT-4o‡ (Hurst et al., 2024)	–	71.9	77.2	66.7	34.7	64.6	66.7
Gemini 1.5 Pro‡ (Team et al., 2024)	–	75.0	81.3	64.4	33.1	74.3	65.8
Qwen3.8 (Qwen Team, 2026)	27B	73.0	76.0	66.7	45.1	70.7	81.0
Open-Source Video MLLMs							
LongVA (Zhang et al., 2024)	7B	52.8	57.6	52.5	37.5	56.0	56.0
Video-LLaVA (Lin et al., 2024)	7B	38.5	43.4	39.0	28.5	47.2	37.4
VideoChat-R1 (Li et al., 2025)	7B	50.4	58.2	49.2	23.8	58.7	65.0
Video-Thinker (Wang et al., 2025)	7B	61.9	65.3	56.0	–	65.2	64.5
Video-R1 (Feng et al., 2026a)	7B	57.6	66.0	57.4	36.9	61.6	61.3
VideoRFT (Wang et al., 2026c)	7B	58.5	65.6	55.1	38.0	44.9	42.7
Time-R1 (Wang et al., 2026e)	7B	60.9	65.1	57.6	37.3	62.9	64.0
ReWatch-R1 (Zhang et al., 2026a)	7B	58.8	65.0	53.6	38.5	60.1	59.8
Qwen2.5-VL-Instruct (Bai et al., 2025b)	7B	55.7	64.5	46.4	32.2	47.8	65.4
Agentic Video MLLMs							
Conan (Ouyang et al., 2026)	7B	55.5	62.8	54.5	38.2	59.2	64.0
VideoMind (Liu et al., 2026d)	7B	58.7	62.3	55.4	41.4	60.8	63.5
LongVT (Yang et al., 2026a)	7B	56.4	59.9	48.5	37.2	56.4	62.2
SAGE (Jain et al., 2026)	7B	44.1	52.4	37.4	31.8	49.7	55.7
VideoZoomer (Ding et al., 2026)	7B	45.3	48.3	39.6	22.9	46.2	61.6
Video-o3 (Zeng et al., 2026)	7B	64.3	66.6	57.2	41.5	64.4	**69.0**
VideoTemp-o3 (Liu et al., 2026c)	7B	57.9	61.3	54.7	36.0	61.8	64.8
Qwen3-VL-Instruct (Bai et al., 2025a)	8B	59.9	68.4	52.2	33.1	58.3	68.0
ParaVT (Yang et al., 2026b)	8B	62.9	**69.9**	56.5	37.3	63.3	64.2
Ours (WORTH)	8B	**66.2**	68.4	**60.8**	**41.6**	**69.3**	66.4

Table 3: **Quantitative Results on Three Temporal Grounding Benchmarks.** All the results are reported in percentage. In this table, the default maximum number of frames fed into each solution is 64. The **Best** and Second Best scores are highlighted in **Bold** and Underline.

Models	Param.	Charades-TimeLens		ActivityNet-TimeLens		QVHighlight-TimeLens	
		R1@0.5↑	*mIoU*↑	*R1@0.5*↑	*mIoU*↑	*R1@0.5*↑	*mIoU*↑
Proprietary Models							
GPT-4o‡ (Hurst et al., 2024)	–	44.5	41.8	41.4	40.4	54.8	52.1
Gemini-2.5-pro‡ (Comanici et al., 2025)	–	61.1	52.8	64.2	58.1	75.9	70.4
Time-R1 (Wang et al., 2026e)	7B	33.7	37.9	28.4	30.3	44.7	42.7
LongVT (Yang et al., 2026a)	7B	16.3	20.4	14.2	16.7	13.8	17.0
Video-o3 (Zeng et al., 2026)	7B	38.6	40.8	26.7	27.4	29.3	30.1
VideoTemp-o3 (Liu et al., 2026c)	7B	32.8	36.3	28.5	28.9	42.5	41.3
Qwen2.5-VL (Bai et al., 2025b)	7B	33.2	35.8	31.9	31.6	41.8	40.2
Qwen3-VL (Bai et al., 2025a)	8B	44.9	45.0	39.8	38.3	54.2	49.7
TimeLens (Zhang et al., 2026b)	8B	**57.2**	50.8	47.0	43.8	61.0	55.2
TimeLens2 (Zhu et al., 2026)	8B	**57.2**	**51.2**	49.6	46.6	65.0	59.3
ParaVT (Yang et al., 2026b)	8B	27.7	33.1	17.9	21.3	24.1	26.8
Ours (WORTH)	8B	52.0	47.6	**51.7**	**47.4**	**67.4**	**61.5**

shares the base model and the frame budget but acquires evidence on every question. WORTH exceeds it by 2.2–6.0% on the five columns without subtitles and trails it only with subtitles, a setting in which ParaVT receives the complete transcript rather than the subtitles of the sampled frames. Against Video-o3, the strongest public agent, WORTH is ahead on five columns and behind only on MMVU, whose questions test domain knowledge rather than localization within the video. Both agents acquire evidence on nearly every question (Section 2.2), and the margins show that this acquisition does not translate into accuracy, as Finding 2 predicts.

Table 4: **Ablations.** All RL rows start from the same cold start and use the same data and interface; only the training signal differs. Accuracy in %, calls as the share of questions with at least one request, tokens as the total per question in units of the overview's token count. The **Best** and Second Best scores of each block are highlighted in **Bold** and Underline.

Setting	VideoMME		LongVideo-Bench	LVBench	MLVU	Charades-TimeLens		Calls	Tokens
	w/o sub	w/ sub				*R1@0.5*	*mIoU*		
(A) Training stage and interface									
Qwen3-VL-8B	59.9	68.4	52.2	33.1	58.3	44.9	45.0	0	1.0
+ direct-action SFT	64.1	**69.0**	59.8	39.0	63.6	51.4	47.3	57	2.9
+ direct-action SFT + GRPO	64.3	67.0	60.1	41.2	65.3	51.9	47.6	92	3.9
+ direct-action SFT + WORTH (**Ours**)	**66.2**	68.4	**60.8**	**41.6**	**69.3**	**52.0**	**47.6**	26	1.3
(B) Credits, on the direct-action cold start									
outcome advantage only (= GRPO)	64.3	67.0	60.1	41.2	65.3	51.9	47.6	92	3.9
+ acquisition credit	65.6	68.0	60.4	41.3	68.4	51.9	47.5	31	1.5
+ observation credit	64.7	67.3	60.6	41.5	66.8	52.0	47.5	90	3.7
+ both (**Ours**)	**66.2**	**68.4**	**60.8**	**41.6**	**69.3**	**52.0**	**47.6**	26	1.3
plan-fixed baseline $\widehat{v}(\varnothing)$ in place of Q_0^μ	64.6	67.2	60.3	41.4	66.0	51.9	47.5	88	3.6

Temporal Grounding. On temporal grounding, WORTH reaches the level of models trained specifically for this task. It improves the base model by 7.1–13.2% in R1@0.5 and 2.6–11.8% in mIoU (Table 3), achieves SOTA performance on ActivityNet and QVHighlights, and matches TimeLens2 within 0.3% on average over the three subsets. It trails TimeLens and TimeLens2 only on Charades, whose 30-second clips the overview already covers, so that acquisition cannot add information there. Every public agent, including ParaVT on the same base model, falls below the base model on all three subsets. Outcome-based training on question answering thus erodes localization, whereas WORTH improves it, since the accepted trajectories of the cold start must cover the annotated span and RL rewards the span directly.

4.3 Ablation Study

Training Stages. The cold start teaches the interface but not when to use it, and outcome-based RL leaves this decision unrepaired. In block (A) of Table 4, the cold start lifts the base model by 5.8% on average but acquires evidence on 57% of the questions, and GRPO adds a further 1.1% only by raising the acquisition rate to 92% and the evidence cost to 3.9 overviews per question. On Video-MME with subtitles, where the transcript settles most questions, these additional reads even lower the accuracy below the cold start, as Finding 1 predicts. WORTH trained from the same cold start improves on GRPO by 1.8% while acquiring on 26% of the questions at one third of the evidence cost.

Acquisition and Observation Credits. The two credits act on different decisions, and their gains are complementary. In block (B), the acquisition credit alone reduces the acquisition rate from 92% to 30% and recovers most of the gain on Video-MME and MLVU, whose questions the overview settles, whereas the observation credit alone keeps the acquisition rate of GRPO and improves LongVideoBench and LVBench, whose plans return several observations. Combining both credits attains the full improvement of 1.8% at the cost of the first, so the acquisition credit governs whether to acquire and the observation credit governs what to request, as the three terms of Eq. (3) predict. Finally, replacing the answering-immediately baseline Q_0^μ with the plan-fixed baseline $\widehat{v}(\varnothing)$ keeps the acquisition rate at 88% and retains only 0.4% of the gain.

5 Conclusion

A long-video agent takes two kinds of decisions, whether to acquire evidence and what to request, and a trajectory-level reward answers neither of the questions that would evaluate them. Our diagnostics of four public agents show the consequences: acquisition is habitual and mostly unearned, most of what it changes precedes the evidence, and almost all returned reports are credited for an outcome they do not affect. WORTH makes both decisions explicit actions of one policy, so that the answer-now alternative and every subset of a plan's reports can be read out from the policy itself, and it supervises each decision with its own comparison: paired continuations from the pre-plan state for acquiring, and exact Shapley values over the reports of a fixed plan for the requests.

ACKNOWLEDGMENTS

AI USE STATEMENT

During the preparation of this manuscript, the authors utilized large language models (specifically, GPT-6) exclusively for language polishing, grammar checking, and formatting assistance. All core scientific ideas, experimental designs, algorithm implementations, and data analyses are conducted solely by the human authors. The authors take full responsibility for the accuracy, integrity, and originality of the content presented in this paper.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the full reproducibility of our research.

Code and Models. Upon acceptance, we will publicly release the complete source code, including the direct-action interface, the evidence-verified cold-start data curation pipeline, and the WORTH reinforcement learning training scripts. An anonymized version of the code and the trained checkpoints is provided in the supplementary material for the review process.

Data and Corpus. The cold-start corpus and the RL training pool are constructed from publicly available datasets. To facilitate exact reproduction, we will release the processed trajectory datasets and the detailed data filtering scripts.

Experimental Details. All architectural details, hyper-parameters, and training configurations are explicitly documented in Appendix D.2 and Appendix D.3.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence, 2026.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

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

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, 2025b.

Leonard Bärmann and Alex Waibel. Where did i leave my keys?—episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1559–1567. IEEE, 2022.

Dan Ben Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Herbench: A benchmark for multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4505–4514, 2026.

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 *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 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, 2026a.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026b.

Mingkuan Feng, Jinyang Wu, Hao Gu, Fangrui Lv, Ruihan Jin, Chuyuan Zhang, Zhengqi Wen, and Jianhua Tao. Taco: Tool-augmented credit optimization for agentic tool use. *arXiv preprint arXiv:2606.30251*, 2026c.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. Treerl: Llm reinforcement learning with on-policy tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12355–12369, 2025.

Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966.

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.

Jitesh Jain, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, Humphrey Shi, et al. Sage: Training smart any-horizon agents for long video reasoning with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41478–41488, 2026.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7969–7992, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

Guankai Li, Jiabin Chen, Yi Xu, Xichen Zhang, and Yuan Lu. Hypereyes: Dual-grained efficiency-aware reinforcement learning for parallel multimodal search agents, 2026a.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*, 2026b.

Mingxuan Li, Kai-Zhan Lee, and Elias Bareinboim. Counterfactual shapley credit assignment, 2026c.

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.

Yanming Li, Xuelin Zhang, WenJie Lu, Ziye Tang, Maodong Wu, Haotian Luo, Tongtong Wu, Zijie Peng, Hongze Mi, Yibo Feng, Naiqiang Tan, Chao Huang, Lian Peng, and Li Shen. Who deserves the reward? sharp: Shapley credit-based optimization for multi-agent system, 2026d.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei, and Wenwu Ou. Ig-search: Step-level information gain rewards for search-augmented reasoning, 2026.

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 *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416, 2026a.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, et al. Videoauto-r1: Video auto reasoning via thinking once, answering twice. *arXiv preprint arXiv:2601.05175*, 2026b.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, et al. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026c.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024.

Ye Liu, Kevin Qinghong Lin, Chang-Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 57481–57506, 2026d.

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. *arXiv preprint arXiv:2602.01334*, 2026.

Delin Mao, Chenghao Sun, Jingwei Song, Chishui Chen, and Linfeng Zhang. Toolvision: Learning when and how to use visual tools with capability-aligned supervision. *arXiv preprint arXiv:2608.08907*, 2026.

Renwei Meng. Policy-conditioned counterfactual credit for verifiable reinforcement learning of long-horizon language agents, 2026.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41089–41099, 2026.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. Videoseal: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. *arXiv preprint arXiv:2605.12571*, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026.

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. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Jiahao Shao, Yuanbo Yang, Yiyi Liao, Yujun Shen, Ceyuan Yang, and Yinghao Xu. Thinking with tools, not with pixels: Tool calls as text scaffolds for visual reasoning, 2026.

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

Lloyd S Shapley et al. A value for n-person games. 1953.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Why pay more when you can pay less: A joint learning framework for active feature acquisition and classification. *arXiv preprint arXiv:1709.05964*, 2017.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Chaoyang Wang, Kaituo Feng, Dongyang Chen, Zhongyu Wang, Zhixun Li, Sicheng Gao, Meng Meng, Xu Zhou, Manyuan Zhang, Yuzhang Shang, et al. Adatooler-v: Adaptive tool-use for images and videos. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18058–18072, 2026a.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, et al. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, volume 2026, pp. 104502–104529, 2026b.

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, 2026c.

Qixun Wang, Yang Shi, Letian Cheng, Zhuoran Zhang, Yan He, Yuqi Tang, Qi Zhang, Xinlei Yu, Ruizhe Chen, Tianrun Xu, et al. Beacon: Knowing when and how to perform agentic visual reasoning. *arXiv preprint arXiv:2607.28595*, 2026d.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

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, 2026e.

Zi-Han Wang, Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Jie Wu, Zhengzhou Cai, Yueqing Sun, Ziang Ye, Linji Hao, Qi Gu, et al. Agentopsd: Recursive self-distillation for agentic reinforcement learning. *arXiv preprint arXiv:2608.05987*, 2026f.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026g.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. Tips: Turn-level information-potential reward shaping for search-augmented llms, 2026.

Shilin Yan, Jintao Tong, Hongwei Xue, Xiaojun Tang, Yangyang Wang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. Act wisely: Cultivating meta-cognitive tool use in agentic multimodal models. *arXiv preprint arXiv:2604.08545*, 2026.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026a.

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. *arXiv preprint arXiv:2605.20342*, 2026b.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. In *International Conference on Learning Representations*, volume 2026, pp. 140958–140985, 2026a.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. Adaptthink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730, 2025.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026b.

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.

Shitian Zhao, Shaoheng Lin, Ming Li, Haoquan Zhang, Wenshuo Peng, Kaipeng Zhang, and Chen Wei. Pyvision-rl: Forging open agentic vision models via rl. *arXiv preprint arXiv:2602.20739*, 2026.

Binbin Zheng, Zijun Xie, Guanqun Zhao, Enlei Gong, Xing Ma, Xiaoliang Fu, and Zeyu Chen. Group-reflective self-distillation for agentic reinforcement learning, 2026a.

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, 2026b.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, et al. Timelens2: Generalist video temporal grounding with multimodal llms. *arXiv preprint arXiv:2607.17423*, 2026.

Wenhao Zou, Xianglong Liu, Wendong Bi, Hanjie Wang, Simin Zhao, and Gong Zhi. Cobra: Learning tool-use boundaries via counterfactual margins. *arXiv preprint arXiv:2609.00967*, 2026.

APPENDIX

The appendix is organized as follows.

- **Appendix A** reviews three lines of related work, namely agentic video understanding, adaptive tool use, and credit assignment for agentic reinforcement learning.
- **Appendix B** shows more details of the diagnostic analysis and explains how the diagnostics motivate the design of WORTH.
- **Appendix C** describes the data sources, the evidence-verified synthesis pipeline of the cold-start corpus, the construction of the RL question pool, and the evaluation benchmarks.
- **Appendix D** specifies the interaction protocol and the training and evaluation details.
- **Appendix E** lists the prompts of the plan, acquire, and assemble stages of data curation.
- **Appendix F** presents qualitative rollouts of the trained policy, including direct answers, single and parallel requests, and multi-plan episodes.

A RELATED WORK

Agentic Video Understanding. Reinforcement learning has advanced long video understanding in multimodal large language models (Feng et al., 2026a; Wang et al., 2026e). However, static sparse sampling inevitably leads to irreversible information loss, motivating agentic systems that can actively determine what to observe. Early systems adopt a training-free paradigm, delegating control to a prompted language model that orchestrates frozen perception modules. VideoAgent (Wang et al., 2024) gathers evidence iteratively, Active Video Perception (Wang et al., 2026g) executes a plan–observe–reflect loop with an explicit sufficiency check, and LensWalk (Li et al., 2026b) plans both the temporal scope and the sampling density of each observation. Subsequent systems train part of the pipeline. VideoMind (Liu et al., 2026d) decomposes the task into role-specific adapters for planning, grounding, verification, and answering, LongVideoAgent (Liu et al., 2026a) trains a master policy that coordinates grounding and vision agents, and VideoSEAL (Qiu et al., 2026) decouples a learned planner from a frozen inspector, attributing ungrounded answers to outcome-only optimization. Recent works embed the perceptual operations in the trajectory of a single model and train it end to end with reinforcement learning. LongVT (Yang et al., 2026a) interleaves global inspection with temporal cropping, SAGE (Jain et al., 2026) switches between single-turn and multi-turn reasoning, Conan (Ouyang et al., 2026) retrieves evidence frames and zooms into segments, and Video-o3 (Zeng et al., 2026) learns interleaved clue seeking with adaptive termination. Furthermore, VideoTemp-o3 (Liu et al., 2026c) couples temporal grounding with question answering, and ParaVT (Yang et al., 2026b) issues multiple temporal requests within a single turn. Despite this architectural diversity, all of these agents are optimized with a single trajectory-level reward, in which answer correctness is augmented by auxiliary terms such as format and tool-call bonuses, as well as temporal overlap with annotated clues. Such terms reward static properties of individual calls rather than their causal effect on the final answer, and the resulting scalar is nonetheless assigned uniformly to every decision along the trajectory.

Adaptive and Efficient Tool Use. Deciding whether to acquire an observation is a classical problem. The value of information (Howard, 1966) prices an observation by the expected improvement of the decision it informs, and active feature acquisition (Shim et al., 2017) learns which features to acquire jointly with the classifier that consumes them. For language models, Toolformer (Schick et al., 2023) retains an API call only if it reduces the loss on the subsequent tokens, whereas FLARE (Jiang et al., 2023) and Self-RAG (Asai et al., 2024) trigger retrieval by low-confidence tokens and by learned reflection tokens, respectively. Similarly, AdaptThink (Zhang et al., 2025) and VideoAuto-R1 (Liu et al., 2026b) decide whether to deliberate before answering. Most of these triggers estimate the need for a tool from the uncertainty of the model rather than from the measured benefit of the tool. For visual tools, the primary difficulty is to sustain an initially weak skill. DeepEyes (Zheng et al., 2026b) incentivizes image cropping through reinforcement learning alone, whereas PyVision-RL (Zhao et al., 2026) and ParaVT (Yang et al., 2026b) counter interaction collapse with accumulative tool rewards. Such incentives sustain tool use but reward the call itself, which is precisely the behavior diagnosed in Section 2.2. Recent work therefore conditions the reward on whether a call is necessary. AdaTooler-V (Wang et al., 2026a) and ToolVision (Mao et al., 2026) rescale the reward by the gap

between answering with and without tools, measured on an external model or on the learner before training, and Beacon (Wang et al., 2026d) determines the reward of tool use by whether the tool-free rollouts of the same group already succeed. Yan et al. (2026) optimize accuracy and efficiency in separate advantage channels, HyperEyes (Li et al., 2026a) adds a reference-adaptive cost for parallel search, and the concurrent CoBRA (Zou et al., 2026) estimates the reward margin between paired trajectories with and without retrieval and optimizes the tool-use boundary with counterfactual marginal advantages. Two diagnostic studies support the concern that such rewards credit the call rather its effect. Ma et al. (2026) find that the gains of vision tool-use reinforcement learning stem mostly from intrinsic learning, and TextCall (Shao et al., 2026) shows that the text of a call explains much of the gain even when the returned image is withheld.

Credit Assignment for Agentic Reinforcement Learning. Search agents such as Search-R1 (Jin et al., 2025) interleave reasoning with retrieval and are trained with outcome rewards under PPO or GRPO (Shao et al., 2024), so that every token of a trajectory receives the same advantage, as in the video agents above. Classical remedies construct a counterfactual baseline by marginalizing the action of one agent (Foerster et al., 2018), redistribute a delayed reward over the steps that caused it (Arjona-Medina et al., 2019), or allocate a joint gain among cooperating players by the Shapley value (Shapley et al., 1953). For language agents, one family rewards a turn by the information it adds. IGPO (Wang et al., 2026b) uses the increase in the probability that the policy assigns to the gold answer, and TIPS (Xie et al., 2026) uses the likelihood under a teacher as a shaping potential. IG-Search (Liang et al., 2026) rewards the query tokens with the gain of the retrieved documents over random ones, and TACO (Feng et al., 2026c) uses the signed difference between answer probes with and without a tool call. These information-gain rewards compare the answers before and after an observation whose request has already been issued. In the terms of Eq. (1), they estimate the effect of the evidence δ_{evi} but not the effect of the request, and they treat one call at a time. A second family localizes the outcome without an explicit intervention, through recurring states or tree branching (Feng et al., 2026b; Hou et al., 2025) or through hindsight self-distillation (Wang et al., 2026f; Zheng et al., 2026a), which leaves implicit the alternative against which a decision is compared. CVT-RL (Meng, 2026) estimates policy-conditioned contributions under deletion and substitution interventions, SHARP (Li et al., 2026d) assigns Shapley-based credit to the agents of a planner–worker system, and counterfactual Shapley credit assignment (Li et al., 2026c) redistributes the reward over time steps.

B Details of the Diagnostic Analysis

B.1 Probing Protocol

We diagnose Video-o3, VideoTemp-o3, LongVT, and ParaVT using their official weights, system prompts, and tool definitions across the five understanding benchmarks detailed in Appendix C.4. To ensure consistency, all agents are provided with an initial video overview comprising at most 64 frames sampled at 1 fps. For each question where an agent initiates a tool call, we freeze the state h (defined in Section 2.1) immediately prior to its first planning step. To instantiate the components a_A, a_B, and a_C from Eq. (1), we complete the trajectory from the frozen state h under three distinct readout conditions using the agent's own weights (Figure B.1):

- **Readout A (Direct Answer):** The agent is forced to generate a final answer directly from h without any further tool interactions.
- **Readout B (Blind Execution):** The agent keeps its planned requests in context, but every observations is replaced with the placeholder "`Observation withheld.`". This allows the agent to process its own requests while being deprived of actual evidence.
- **Readout C (Full Execution):** The agent executes the planned tool calls and receives the actual returned observations.

Across all readouts, we employ greedy decoding and constrain the output generation using a grammar that strictly enforces a single-letter final answer. For ParaVT, which generates multiple tool calls in its initial planning step, we additionally evaluate performance over all possible subsets of the returned observations to measure the marginal utility of each piece of evidence.

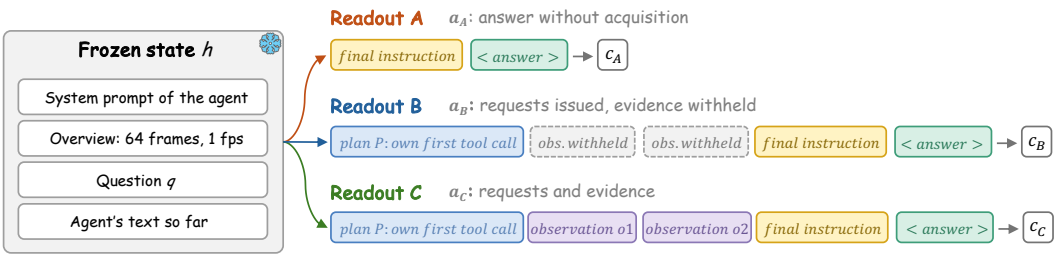


Figure B.1: **Illustration of the probing protocol.** For a given frozen state h prior to the agent's first tool-calling plan, we evaluate three readout conditions (A, B, and C) to separate the effect of the requests (a_B against a_A) from the effect of the returned evidence (a_C against a_B).

Table B.1: **Reward design of the diagnosed agents.** All reward terms are summed into a single trajectory-level scalar. r_a and r_f denote the answer reward and format reward, respectively.

Agent	Reward	Required signal	Algorithm
Video-o3	$R = r_a(1+\beta) + r_f$, where $\beta = (b_0 + w_g \mathcal{S}_{\text{clue}})\gamma$ and γ decays after a reference turn budget.	Clue spans	DAPO
VideoTemp-o3	$R = r_a + r_f + r_{\text{IoU}}$, with a penalty of 0.1 is applied when $\text{IoU} < 0.1$.	Annotated spans	GRPO
LongVT	$R = R_{\text{acc}} + R_{\text{format}} + R_{\text{time}}$, where $R_{\text{acc}} \in \{1, 0.5, 0\}$ and R_{time} is computed from span IoU.	Spans + LLM judge	GRPO
ParaVT	$R = r_a + r_f + r_{\text{tool}}$, where r_f includes an exploration-anchoring bonus and r_{tool} assigns 0.1 for a parseable call, 0.3 for an executed call, and $+0.5$ if the answer is correct.	None	GRPO

Every readout is deterministic, so the uncertainty of a statistic stems from the sample of videos on which it is measured. We therefore report 95% percentile intervals from 10,000 bootstrap resamples that draw videos with replacement within the categories of each benchmark and carry all questions of a video together with their three readouts. Resampling videos accounts for the dependence among the questions of one video, and keeping the three readouts together yields the interval of a paired difference such as δ_{acq}. All probing experiments are conducted on a cluster of eight NVIDIA A100 GPUs. The full evaluation protocol requires approximately 13 hours of total execution.

B.2 Rewards of the Diagnosed Agents

None of the diagnosed agents is rewarded for the effect of an acquisition (Table B.1). Every agent optimizes a single trajectory-level scalar with a group-relative algorithm, and the auxiliary terms score the format, the act of calling, the overlap of a window with an annotated span, or the number of turns. No term compares a trajectory with answering immediately from the same state, so nothing in the training signal separates a necessary acquisition from a habitual one.

B.3 Contrast Collapse under a Group-Relative Advantage

The supervision that a group-relative advantage provides on the decision to acquire is proportional to $p \cdot (1-p)$ and vanishes at the measured acquisition rates. Consider a state where the agent acquires evidence with probability p. Let Q_1 and Q_0 denote the expected rewards for acquiring and answering immediately, so that $\Delta = Q_1 - Q_0$. As the group size grows, the expected group mean reward converges to $pQ_1 + (1-p)Q_0$. Subtracting this baseline yields an expected mean-centered advantage of $(1-p)\Delta$ for an acquiring rollout and $-p\Delta$ for an answering one. While these advantages carry the correct signs, computing the expected gradient with respect to the acquisition logit z requires weighting each branch by its occurrence probability and its policy score $\partial_z \log \mu$. Since the score equals $1-p$ for acquiring and $-p$ for answering, the expected gradient evaluates to

$$\mathbb{E}\big[A\,\partial_z \log \mu\big] = p(1-p)\Delta(1-p) + (1-p)(-p\Delta)(-p) = p(1-p)\Delta. \tag{8}$$

Normalizing this quantity by the standard deviation of the group rewards merely rescales the gradient without altering its sign.

In practice, a finite group provides no information about Δ if all its rollouts select the same initial action. In such homogeneous groups, the empirical advantages are identically zero, yielding a null gradient. For a group of $G = 8$ rollouts, the probability of sampling a mixed group containing both actions is $1 - p^8 - (1 - p)^8$ (Figure B.2). Models like LongVT and ParaVT operate on the flat right tail of this distribution, where at most 1.2% of groups are mixed and the gradient signal from Eq. (8) retains merely 0.6% of its peak value. Conversely, Video-o3 resides on the descending flank, leaving VideoTemp-o3 as the only model that consistently maintains mixed groups across most benchmarks.

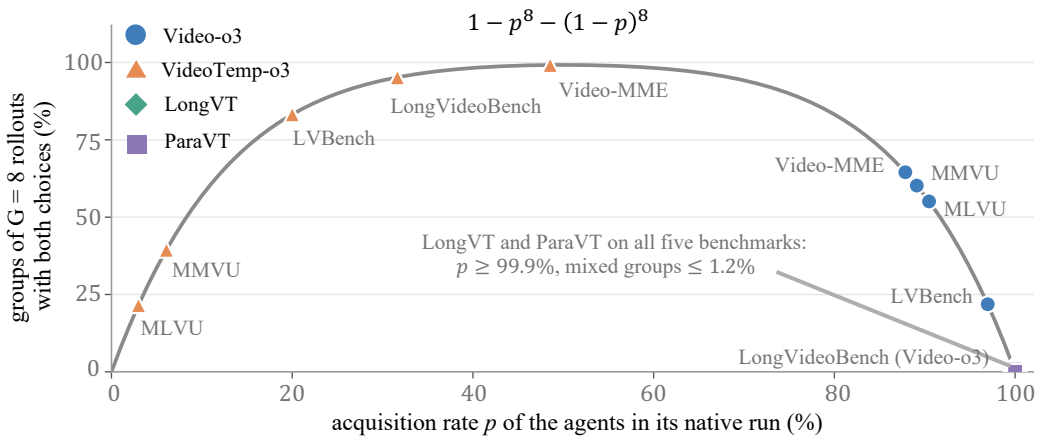


Figure B.2: **Contrast collapse.** Probability that a group of $G = 8$ rollouts contains both an acquiring and an answering rollout as a function of the acquisition rate p. Each marker is one agent on one benchmark at its native rate.

B.4 Answer Revisions

A revision made after observing the evidence is no more accurate than one made with the evidence withheld. Net effects average over helped and harmed answers and therefore do not indicate whether a change carried information. We hence examine each revision at the level of the option letter, on the states whose three readouts are valid (7,723 / 2,137 / 8,379 / 8,365 states with 4.2 options on average). An answer is revised at the request stage when $a_B \neq a_A$ and at the evidence stage when $a_C \neq a_B$, and Figure B.3a reports how often a revised wrong answer lands on the correct option. The observed 47–53% exceeds the 31–32% of an undirected re-draw among the remaining options, indicating a move to an option that the agent already ranked highly. However, the difference between the evidence stage and the request stage lies within $\pm 1.6\%$ for every agent and is never distinguishable from zero. The revisions in a_B are not artifacts of the withheld context either, since 41–78% of them persist once the evidence is shown (Figure B.3b).

B.5 Observation-Level Attribution

Two facts of the 7,750 parallel plans of ParaVT underlie the requirement of Section 2.4 that an observation credit be signed and averaged over subsets. First, 361 of the 3,027 plans whose readout is wrong contain a subset of their own observations from which the readout is correct, so that a credit restricted to rewarding observations cannot suppress the requests that mislead the answer. Second, among the 189 plans whose full set of observations turns a wrong readout into a correct one, 125 contain at least two observations that are individually sufficient, so that any credit computed from single removals assigns zero to all of them.

B.6 The Value of Acquisition Is Policy-Specific

Whether an acquisition pays depends on the policy that acquires, which is why the first requirement of Section 2.5 estimates this value for the current policy rather than transferring it from another

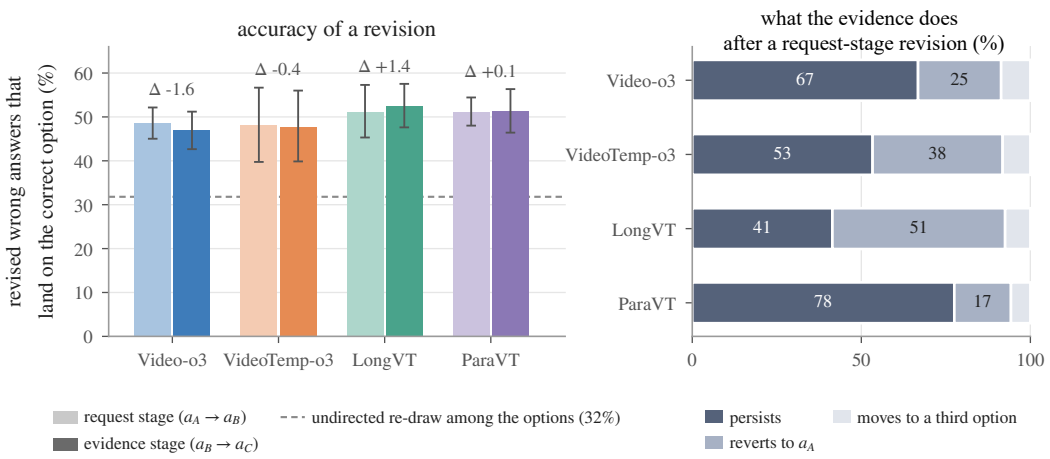


Figure B.3: **Answer revisions. (a)** Share of revised wrong answers that land on the correct option at the request stage and at the evidence stage, with the difference Δ between the stages and the accuracy of an undirected re-draw as a dashed line. **(b)** Fate of a request-stage revision once the evidence is shown.

model. We group the questions of each agent and benchmark by the categories of the benchmark, keep the groups with at least 80 questions, and compute the net effect δ_{acq} of each group (Figure B.4). Of the 102 agent–category cells, 10 have an interval that excludes zero against 5.1 expected by chance, so the effect is structured rather than noise around zero. The structure does not transfer across agents. ParaVT gains $+6.8\%$ $[3.4, 10.8]$ on the egocentric questions of MLVU and loses 10.2% $[-16.5, -4.4]$ on its counting questions, whereas Video-o3 and LongVT show no distinguishable effect on either category. At the level of individual questions, the sets of questions that acquisition helps overlap across agents only 0.5–1.8 times as often as independence predicts. The value is nonetheless substantial once it is measured for the right policy. An oracle that chose per question between a_A and a_C would exceed the better of the two uniform choices by 1.8–5.6%, an upper bound because part of the gap reflects the request-induced perturbations of Section 2.3.

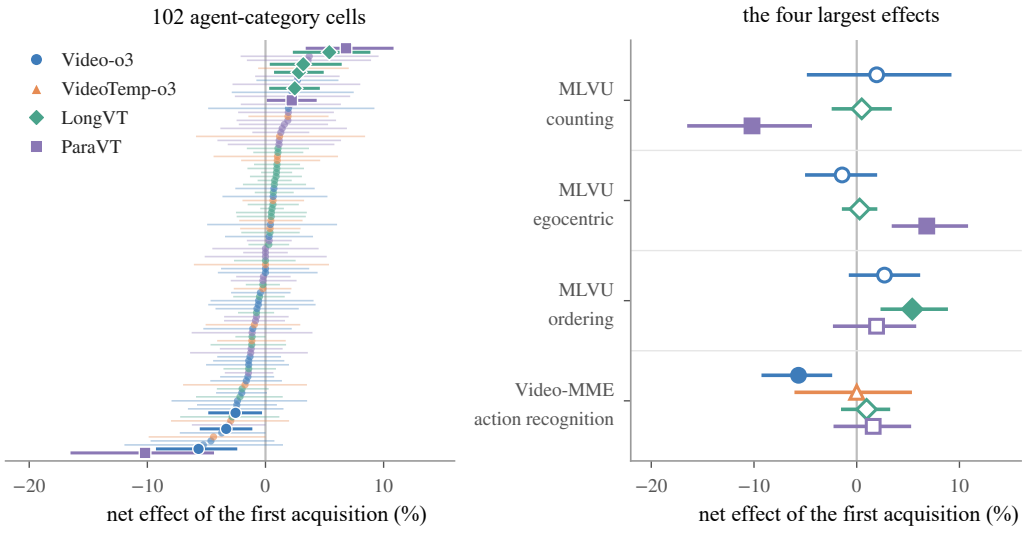


Figure B.4: **Policy-specific value of acquisition. (a)** Net effect of the first acquisition in the 102 agent–category cells with at least 80 questions, sorted by effect, with 95% intervals and filled markers for the 10 intervals that exclude zero. **(b)** The four largest effects, with the same categories for the other agents.

B.7 From the Diagnostics to the Design of WORTH

Two choices of the interface in Section 3.1 follow from the diagnostics.

Direct Action. Both credits compare continuations of the policy that differ in a single decision. Under the direct-action interface, an assistant turn is either a plan or an answer, so that the first token of a turn commits to the decision and answering immediately is a continuation with a forced first token. The branch prefix that receives the acquisition credit is thus a short and well-defined span, whereas a reasoning trace in front of the action would spread the credit over tokens that also serve the answer. Section 2.3 adds a second reason. The text written before any evidence arrives already perturbs the answer, so that a long trace in front of a plan would entangle the value of the plan with the effect of writing it.

Textual Observations. The observation credit requires observations that can be withheld individually and answers that can be read out for every subset. A textual observation of at most 1,024 tokens keeps both inexpensive, because the 2^K readouts of a plan are short answers on a cached prefix. The frames of a window would instead cost about 76 tokens each and exceed the overview several times over. Textual observations also expose the evidence cost of Section 3.2, since the frames processed by the reader are counted per request. The price is that an observation is a lossy summary, so the observation credit measures the value of an observation as delivered, including the mistakes of the reader.

C Data Curation and Benchmarks

C.1 Data Sources

Both training stages draw on one pool of public video question-answering and temporal-grounding datasets (Figure C.5). The pool comprises ParaVT-Source (Yang et al., 2026b), namely Video-R1 (Feng et al., 2026a), GeminiCoT, LongVideoReason, Selftrace, MuSeg, and TVG, together with CG-Bench (Chen et al., 2025), TimeLens2 (Zhu et al., 2026), Charades-STA, QaEgo4D (Bärmann & Waibel, 2022), TVQA (Lei et al., 2018), YouCook2 (Zhou et al., 2018), MovieChat (Song et al., 2024), ET-Bench (Liu et al., 2024), and HER-Bench (Ben Ami et al., 2026). We keep the questions with a verifiable answer, namely a multiple-choice letter or a temporal span, remove every video that appears in any evaluation benchmark, and cap the number of questions per video so that no source or video dominates. The resulting pool contains 233,246 questions on 54,351 videos.

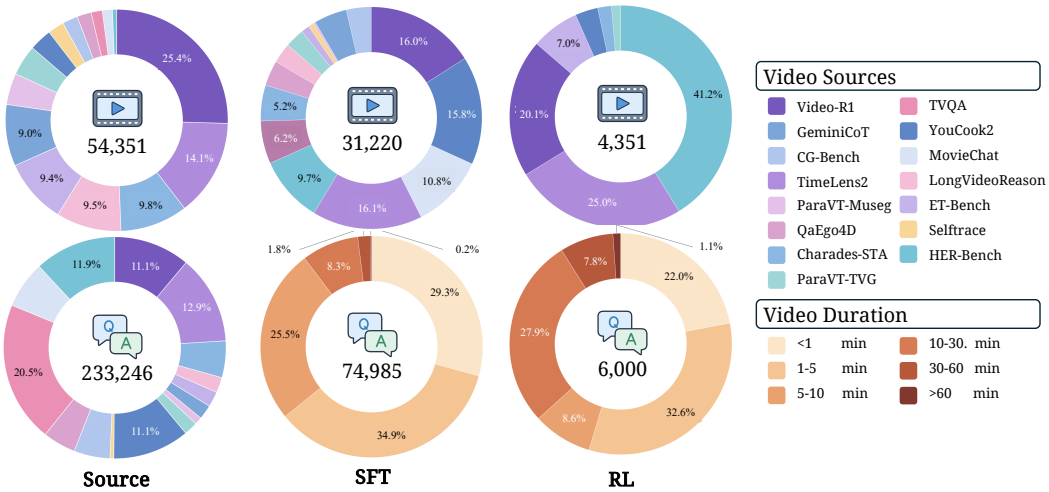


Figure C.5: **Composition of the data.** Video sources (top) and video durations (bottom) of the source pool (left), the cold-start corpus (middle), and the RL question pool (right). The center of each ring gives the number of videos (top) or of questions (bottom).

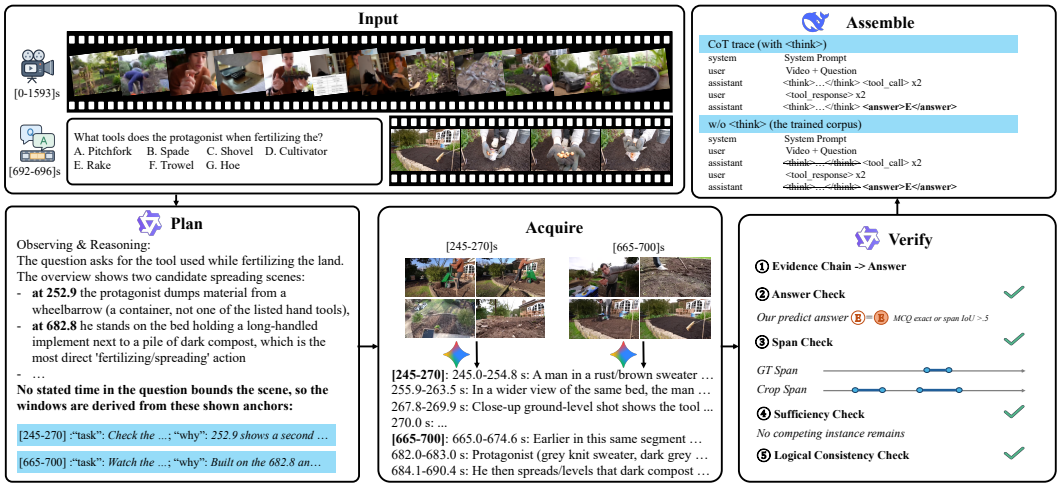


Figure C.6: **Evidence-verified synthesis of the cold-start corpus.** A planner proposes anchored windows from the question and the overview (plan), a reader returns one observation per window (acquire), five checks accept or reject the finished trajectory against the ground truth and an independent audit (verify), and accepted trajectories are serialized with and without a reasoning trace (assemble).

C.2 Evidence-Verified Synthesis of the Cold-Start Corpus

The cold-start corpus is synthesized by a four-stage pipeline, namely plan, acquire, verify, and assemble (Figure C.6). The generator conditions exclusively on the policy inputs, whereas the answer and the annotated evidence span serve solely for trajectory verification. The prompts for the plan, acquire, and assemble stages are detailed in Appendix E.

Plan. A planner first determines whether the overview contains sufficient information to answer the question, prompting a direct response if so. Otherwise, it proposes up to five temporal windows, each anchored to a timestamp presented in the overview or mentioned in the question, along with a specific retrieval objective for each window. If no explicit temporal anchor is available, the planner requests a global scan of the entire video.

Acquire. A reader processes each requested window at 1 fps and returns a single timestamped observation per window, mirroring the interface protocol described in Section 3.1. A subsequent reasoning step either generates an answer from the acquired observations or issues a new plan. Consequently, a trajectory may comprise multiple planning steps, with each decision conditioned strictly on the observations acquired up to that point.

Verify. A trajectory is accepted only if it satisfies the three criteria outlined in Section 3.1, operationalized in the pipeline as the five checks illustrated in Figure C.6. The *answer check* requires the final answer to match the ground truth, using exact matching for multiple-choice questions and a temporal IoU > 0.5 for spans. The *span check* ensures that the requested windows cover the annotated evidence span, provided one exists. The remaining checks are performed by an independent text-only audit, which is provided with the trajectory but not the video or the ground-truth answer. The *evidence-chain check* mandates that every claim in the final answer be supported by an acquired observation, while the *sufficiency check* ensures that no alternative instance of the queried event remains unexamined. The *logical consistency check* verifies that the reasoning steps within the trajectory are self-consistent and free of internal contradictions. Any trajectory failing a check is discarded outright rather than repaired.

Assemble. Accepted trajectories are deduplicated by question, capped per video, mixed by a duration-balanced recipe, and serialized in the interface format. A writer module additionally converts the planner's analysis into a reasoning trace, which is subsequently stripped from the WORTH corpus to ensure that each assistant turn consists exclusively of either a plan or a final answer. The final corpus contains 74,985 trajectories spanning 31,220 videos. Among these, 33% answer directly, 27% contain a single request, 17% issue multiple requests within a single plan, 8% involve two or more plans, and 15% are temporal-grounding answers.

C.3 RL Question Pool

The RL training pool is sourced from the same underlying datasets but is strictly disjoint from the cold-start corpus at the question level, and from all evaluation benchmarks at the video level. Because the group-relative advantage provides no learning signal when all rollouts uniformly succeed or fail, we curate the question pool based on the outcome distribution of the cold-start policy. Specifically, each candidate question is queried four times under the training interface. We retain multiple-choice questions that exhibit mixed outcomes, a targeted subset of questions the policy consistently answers incorrectly, and temporal-grounding questions that demonstrate variance in IoU across the four runs. The resulting pool comprises 6,000 questions (4,000 multiple-choice and 2,000 temporal-grounding).

C.4 Benchmarks

Long-video understanding is evaluated on five multiple-choice benchmarks. Video-MME[1] contains 2,700 questions on 900 videos of short, medium, and long duration and is evaluated with and without subtitles. LongVideoBench[2] contributes the 1,337 questions of its validation split, LVBench[3] 1,549 questions on videos that average more than an hour, MLVU[4] the 2,174 multiple-choice questions of its development split, and MMVU[5] the 625 multiple-choice questions of its validation split. Temporal grounding is evaluated on TimeLens-Bench[6], which re-annotates Charades-STA, ActivityNet Captions, and QVHighlights, by R1@0.5 and mIoU on each subset. The diagnostics of Section 2 use the five understanding benchmarks with Video-MME evaluated without subtitles, 8,385 questions in total, and the ablation averages of Section 4.3 are taken over Video-MME without subtitles, LongVideoBench, LVBench, and MLVU.

D Experiment Details

D.1 Interaction Protocol

The interface exposes one tool, `crop_video(task, start_time, end_time)`, and no reasoning, so that an assistant turn consists either of tool calls or of an answer. A plan is one assistant turn with at most five calls, and an episode allows at most three plans. The overview consists of at most 64 frames sampled at 1 fps, uniformly when the video is longer than 64 seconds, each frame limited to 50,176 pixels and preceded by its timestamp, which amounts to about 6k tokens per question. The reader is the policy itself, with shared weights and no gradient. It executes the requests of a plan in parallel on the original video, reads each window at 1 fps with at most 64 frames at the resolution of the overview, receives the task of the request but not the answer options, and returns one observation of at most 1,024 tokens per request. A window that exceeds the video is clipped, and the observations of a plan are appended to the state as one user turn. The task reward $R \in [0, 1]$ is exact match for multiple-choice questions. For grounding, R is the mean of the temporal IoU and the similarity $\exp(-W_1/|g|)$ between the predicted span and the annotated span g, where W_1 is the Wasserstein distance between the two intervals, so that a prediction that misses the span still receives a graded reward (Zhu et al., 2026). Figure D.7 illustrates one episode.

D.2 Training Details

Cold Start. The cold start fine-tunes Qwen3-VL-8B-Instruct on the evidence-verified corpus of Appendix C.2 together with reader demonstrations (Table D.2). The loss is next-token cross-entropy on the assistant tokens, sequences are packed to 16,384 tokens, and samples above this length are dropped rather than truncated. The video input follows the evaluation protocol exactly, so that the policy receives the same overview at training and at test time.

[1] https://huggingface.co/datasets/lmms-eval/Video-MME
[2] https://huggingface.co/datasets/longvideobench/LongVideoBench
[3] https://huggingface.co/datasets/lmms-eval/LVBench
[4] https://huggingface.co/datasets/MLVU/MVLU
[5] https://huggingface.co/datasets/lmms-eval/MMVU
[6] https://huggingface.co/datasets/TencentARC/TimeLens-Bench

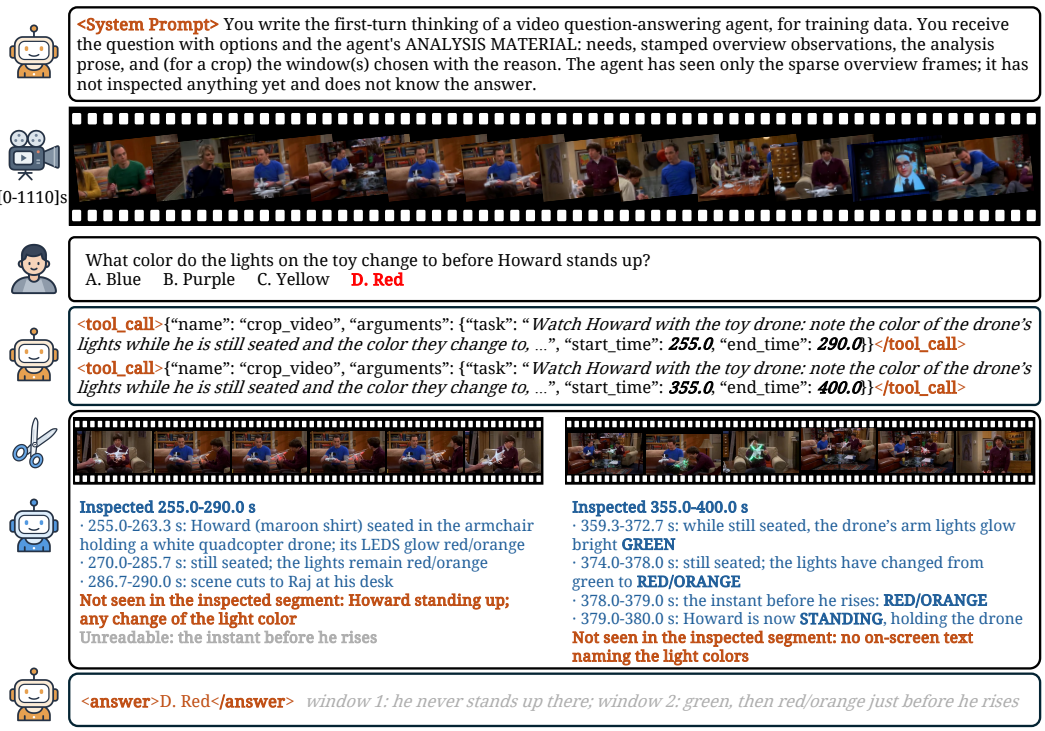


Figure D.7: **One episode under the direct-action interface.**

Table D.2: **Cold-start hyper-parameters.** The video input configuration is identical to the evaluation protocol.

Category / Hyper-parameter	Value / Configuration
Model & Data	
Base model	Qwen3-VL-8B-Instruct
Training corpus	79,210 evidence-verified trajectories, 27,259 reader demonstrations
Optimization	
Training steps / Epochs	2,000 / ≈1
Optimizer / Learning rate	AdamW / 2×10^{-5}
LR schedule	Cosine decay with 5% warm-up
Weight decay / Grad. clip	0.01 / 1.0
Precision	`bf16`
Loss function	Next-token cross-entropy (assistant tokens only)
Input & Sequence Processing	
Overview (Video)	64 frames at 1 fps; 784–50,176 pixels/frame; 1 timestamp/frame
Reader observations	1 window at 1 fps ($\leq$64 frames); 1 observation/window
Max sequence length	16,384 tokens (longer sequences dropped)
Batch size	2 per GPU $\times$ 8 GPUs

Reinforcement Learning. The RL stage starts from the cold start and runs outcome-based GRPO or WORTH with otherwise identical settings (Table D.3). Rollouts are generated asynchronously by two vLLM replicas and consumed by six FSDP trainers, at most two steps behind the current policy. Each step samples 42 questions with $G = 8$ rollouts at temperature 0.7, and one epoch over the 5,915 questions of the pool takes 110 steps. The reward contains no format term, since the interface has no reasoning trace. The tool-bonus arm adds the graded tool reward released with ParaVT, which pays 0.1 for a parseable request, 0.3 once an observation is returned, and 0.5 more when the answer is also correct, capped at 1.

Table D.3: **RL hyper-parameters.** The first block is shared by the GRPO baselines and WORTH, while the second block is specific to WORTH.

Hyper-parameter	Value / Configuration
Shared Setup (GRPO & WORTH)	
Rollouts (G) / Questions per step	8 / 42
Total training steps	140
Sampling temperature	0.7
Optimizer / LR / Weight decay	AdamW / 2×10^{-6} (constant) / 0.01
Grad. clip / PPO minibatches	1.0 / 2
Clip ratio (ϵ) / KL coeff.	0.2 / 0.0
Interface limits	5 requests/plan, 3 plans/episode
Generation limits	4,096 tokens/turn, 19,200/episode
Overview input	64 frames @ 1 fps, $\leq$50,176 px/frame, 1 ts/frame
Reader execution	Policy inference, 1 fps, $\leq$64 frames/win, $\leq$1,024 tokens/obs
Task reward	Exact match (MCQ), $\frac{1}{2}$(IoU + TW) (Grounding)
Tool bonus (if enabled)	+0.1 (request), +0.3 (obs), +0.5 (ans), capped at 1.0
Validation	472 held-out Qs, greedy decoding, every 20 steps
Infrastructure	8 $\times$ A100-80G (2 vLLM, 6 FSDP)
Step time	4.8 m (GRPO), 7.4 m (+Tool), 6.4 m (WORTH)
WORTH-Specific Components	
Min. branch size (m): acq / ans	4 / 8
Readouts per subset (M_v)	4 (max 32/rollout, $M_v = 2$ for $K = 4$)
Readout length / temperature	32 tokens / 0.7
Observation cost weight (λ)	0.01 per reader overview token
Credit normalization	Norm. by group reward dev. (clip 0.1), then RMS over batch
Credit scales (α, β) / clip	1.0, 1.0 / 3.0

D.3 Computation of the Credits

Forced Continuations. The G rollouts of a question are split by their first action. If the acquiring branch has fewer than four members or the answering branch fewer than eight, the missing continuations are generated from the same prompt by forcing the first tokens of the turn, and they are used for scoring only. A forced answer costs a few tokens on the cached prompt, whereas a forced acquisition is a complete scoring-only episode with reader calls and constitutes the only added reader work.

Subset Readouts. For a rollout whose first plan returns $K \leq 5$ observations, every subset of the observations is read out M_v times as an answer-only continuation at the rollout temperature. A withheld observation is replaced by a fixed placeholder at its position, the retained observations keep their content and order, and the reader is not re-run. Repeated readouts of one subset share the cached context, so that M_v costs decoding only, and a budget of 32 readouts per rollout lowers M_v to 2 for $K = 4$ and to 1 for $K = 5$. The full subset reuses the reward of the rollout itself whenever the rollout answered directly after its first plan. The identity $\sum_i \widehat{\phi}_i = \widehat{v}(N) - \widehat{v}(\varnothing)$ holds by construction and serves as a check of the enumeration rather than as evidence that the readouts are noise-free.

Scaling and Objective. Both credits are divided by the reward deviation of their group, floored at 0.1, so that a grounding question with IoU differences of 0.1 receives the same scale as a multiple-choice question with differences of 1, and then by the root mean square of the valid credits of the trainer batch, followed by clipping at 3. The update is the clipped objective of GRPO with the routed advantages of Eq. (7) as detached targets,

$$\mathcal{L}_{\mathrm{W}}(\theta) = \sum_{j,k} w_{j,k}\, \ell\big(\rho_{j,k}(\theta),\, \mathrm{sg}[A^{\mathrm{W}}_{j,k}]\big), \qquad \ell(\rho, A) = -\min\{\rho A,\ \mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)\, A\}, \quad (9)$$

where $\rho_{j,k}$ is the ratio of π_θ to the proximal policy and $w_{j,k}$ the token weights of the trainer. Counterfactual continuations and readouts enter training only through $\widehat{\Delta}$ and $\widehat{\phi}$ and are never added as training trajectories.

D.4 Evaluation Protocol

Comparisons. We compare our method with the followings. Qwen3.8-27B[7], LongVA[8], Video-LLaVA[9], VideoChat-R1[10], Video-Thinker[11], VideoRFT[12], Time-R1[13], ReWatch-R1[14], Qwen2.5-VL-Instruct-7B[15], Conan[16], VideoMind[17], LongVT[18], SAGE[19], VideoZoomer[20], Video-o3[21], VideoTemp-o3[22], Qwen3-VL-Instruct-8B[23], ParaVT[24], TimeLens[25], TimeLens2[26].

Long Video Understanding. Every reproduced baseline runs its released weights, prompts, decoding parameters, and tools, and only the video input is unified to 1 fps and at most 64 frames. The tools of an agent operate on the original video as released. ParaVT runs its released driver, which encodes the overview at native resolution. In the subtitle setting of Video-MME, ParaVT prepends the complete transcript, whereas every other row receives the subtitles aligned with the sampled frames following lmms-eval. MLVU is reported as the macro average over its seven tasks, and MMVU as the accuracy on its 625 multiple-choice questions. The rows of VideoChat-R1, Video-Thinker, Video-R1, VideoRFT, ReWatch-R1, Qwen2.5-VL, Conan, SAGE, and VideoZoomer are the numbers reported by ParaVT under its 64-frame protocol, and the Qwen3-VL-Instruct row is the number ParaVT reports for the base model under its tool prompt.

Temporal Grounding. All rows are evaluated on the same 64-frame overview with exact timestamps. Base models use the official grounder prompt of TimeLens-Bench, grounding models use the prompts of their model cards, and agents use their released tools. R1@0.5 compares the first predicted span with the single annotated span, and an unparseable prediction counts as an IoU of zero and stays in the denominator.

Evidence Cost. For every question we log the tokens that enter and leave the policy over all turns and the tokens that enter and leave the reader over all requests, and we divide their sum by the token count of the first turn, which consists of the overview and the question. Answering immediately therefore costs 1.0, and the cost of a question that acquires evidence includes the observations that the policy re-reads in later turns.

[7] https://github.com/QwenLM/Qwen3.8
[8] https://github.com/evolvinglmms-lab/longva
[9] https://github.com/PKU-YuanGroup/Video-LLaVA
[10] https://github.com/opengvlab/videochat-r1
[11] https://github.com/shijian2001/Video-Thinker
[12] https://github.com/QiWang98/VideoRFT
[13] https://github.com/www-Ye/Time-R1
[14] https://github.com/alibaba/ReWatch-R1
[15] https://huggingface.co/collections/Qwen/qwen25-vl
[16] https://github.com/OuyangKun10/Conan
[17] https://github.com/yeliudev/VideoMind
[18] https://github.com/EvolvingLMMs-Lab/LongVT
[19] https://github.com/allenai/SAGE
[20] https://github.com/zsgvivo/VideoZoomer
[21] https://github.com/MCG-NJU/Video-o3
[22] https://github.com/Kwai-Keye/VideoTemp-o3
[23] https://github.com/qwenlm/qwen3-vl
[24] https://github.com/EvolvingLMMs-Lab/ParaVT
[25] https://github.com/TencentARC/TimeLens
[26] https://github.com/MCG-NJU/TimeLens2

E PROMPT FOR DATA CONSTRUCTION

This section presents the prompts used for data curation in Figure E.8 to Figure E.10. Figure E.8 presents the prompt for the Plan Stage which sees only the timestamped overview and the question. Figure E.9 shows the prompt for the Acquire Stage, in which the reader inspects one window for one task without seeing the options. Figure E.10 illustrates the prompt for the Assemble Stage, in which a writer turns the planner's analysis into the reasoning of the trajectory.

Prior Input :
[Original Video, Question]

Prompt for Plan Stage :

You plan the first move of a long-video question-answering agent. You see a sparse OVERVIEW (at most 64 frames sampled across the whole video, each preceded by its timestamp like <105.0 seconds>) and a multiple-choice question. You do NOT know the answer. Every option is a hypothesis to be tested, never a fact.

Epistemics. The overview is sparse: nothing between two shown timestamps was observed, and "not in the shown frames" never means "not in the video". Long videos may show one frame per several minutes.

Your job is ANALYSIS, not captioning. Captions have no value; what matters is the reasoning from what the question requires, through what the frames show, to what must be inspected:
1. needs — the smallest set of visual facts that must be established to answer (1-4). Do not write one need per option. Typical shape: (a) locate the reference event/person the question describes, (b) read the asked attribute/relation at that moment, (c) if the question says first/again/before/after, establish the order.
2. observations — only frames that bear on a need, with the exact displayed timestamp and the directly visible fact (at most 8). If the question's reference event may occur more than once, check ALL frames for every plausible occurrence and list each. Decoys are common in these videos (a second boutique, a second person in pink, a second fireworks show): when two or three shown frames could each be the reference event, you MUST cover each candidate with its own window instead of picking the most salient one — the close look, not the overview, decides.
3. analysis — prose. What the overview establishes for each need (cite timestamps); what is still missing and WHY that gap prevents choosing one option; whether a shown timestamp or a time stated in the question bounds where the missing fact must be; how you derive the window from that anchor (the activity starts before / continues after the shown frame, scene boundaries, typical duration). This analysis is the material a writer will turn into the agent's reasoning, so make every sentence carry an inference, not a description.
4. route —
DIRECT: every need is SEEN at the precision the options require and the frames leave only one option possible. General knowledge or option wording never makes DIRECT.
CROP: some need is PARTIAL/UNSEEN and the overview gives a reason to look somewhere: a shown frame that IS the reference moment (an anchor), or shown frames that merely COULD be it, or an explicit time in the question, or the video start/end. Propose 1-3 non-overlapping windows, each 8-<<WINDOW_MAX>> seconds, endpoints at multiples of 5 seconds or exact shown timestamps; each window must contain the shown timestamp (or question time) it is built on, and "why" must name that timestamp and justify both ends. A window may be a CANDIDATE: a shown frame that could be the moment but is not confirmed (a similar scene, the same person elsewhere, a second occurrence) — say so in "why" and give EVERY such candidate its own window, because the close look, not the overview, decides which one is the moment. Prefer two or three candidate windows over one confident window whenever the frames leave more than one plausible location; never place a window where no frame or stated time points. When the anchor frame is the first (or last) shown frame of the scene, or shows a sentence or action still in progress, the event most likely began before (or continues after) it: extend the window generously on the unseen side, up to the <<WINDOW_MAX>>-second limit, rather than centring a short window on the frame. When the question states a time ("the 17th minute", "[15:47,15:56]"), the window must cover the whole stated interval (a minute N means all of minute N) plus margin. Give each window a concrete reader task (a sentence or two) asking for the missing visual facts and the identity cues needed to confirm it is the right moment — never asking to confirm an option, never asking to listen.
UNANCHORED: nothing bounds where the missing fact is — no shown frame could be the moment (not even as a candidate), no time is stated, and neither the video start nor its end is a reason to look there. windows=[]; say in "missing" what could not be anchored. Use it only after looking for candidate frames: a plausible candidate window beats UNANCHORED.
AUDIO: the decisive fact can only come from speech, narration or sound (a stated reason, a spoken date, a name said aloud) and no on-screen text could carry it. The reader sees frames only, so say so in "missing" instead of proposing a window.
5. On DIRECT, "choice" is the option letter. For CROP/UNANCHORED/AUDIO, "choice" is null and no field may state or favor an answer.

Return exactly one JSON object:

```
{
 "needs": [{"need": "...", "status": "SEEN|PARTIAL|UNSEEN", "seen_at": [105.0], "missing": null}],
 "observations": [{"t": 105.0, "fact": "..."}],
 "analysis": "...",
 "route": "DIRECT|CROP|UNANCHORED|AUDIO",
 "windows": [{"start": 95.0, "end": 125.0, "task": "...", "why": "..."}],
 "choice": null
}
```

No markdown, no text outside the JSON.

Figure E.8: **Prompt for Plan Stage of Data Curation**

Prior Input :
[Video Segment, subtask, Question]

Prompt for Acquire Stage :
You are a segment sub-agent of a long-video question-answering agent. The main agent asked you to inspect ONE segment of the video for a task. You see the segment sampled exactly the way the main agent sees the whole video: at most 64 frames, at most one per second, each preceded by its absolute timestamp like <104.5 seconds>. You do not know the question's options.

Decide what to do:
- REPORT when the frames let you state the task's facts at the needed precision.
- CROP when this segment is longer than 64 seconds (so you saw fewer than one frame per second), the task-relevant moment is visible or bracketed by shown frames, and the decisive detail (text, a small object, a brief action) falls between or is only glimpsed in the shown frames. Then request ONE closer look: a sub-segment inside this segment, 8-64 seconds long, containing the relevant frames, with a sub-task. You will receive its report and then write your final report. Never request a closer look for a segment of 64 seconds or less (you already see every second), and never to hunt for content that is simply absent from this segment.

Report contract — the main agent will reason from it, so write facts that someone who has not seen the frames can use:
- facts: 3-8 entries, each tied to the absolute seconds where they are visible. First the identity cues that confirm this is the right moment (who: appearance, clothing; what action; where: scene; on-screen text in quotes), then the task's target attribute at full precision (exact colour, exact text, left/right from the viewer's side, counts), then temporal order.
- absent: things the task asks about that do NOT appear anywhere in this segment — scoped strictly to this segment.
- unreadable: target details too small or blurry to read. Never guess them.

Return one JSON object:
{"decision": "report|crop",
 "report": {"facts": [{"span": [104.5, 106.0], "fact": "..."}], "absent": ["..."], "unreadable": ["..."]},
 "sub_window": {"start": 110.0, "end": 140.0, "task": "..."},
 "why": "..."}
report is null when decision is crop; sub_window is null when decision is report. Only what the frames show; no option guessing; no markdown.

Figure E.9: **Prompt for Acquire Stage of Data Curation**

Prior Input :
[Question, Plan Process, Acquire Process]

Prompt for Assemble Stage :
You write the first-turn thinking of a video question-answering agent, for training data. You receive the question with options and the agent's ANALYSIS MATERIAL: needs, stamped overview observations, the analysis prose, and (for a crop) the window(s) chosen with the reason. The agent has seen only the sparse overview frames; it has not inspected anything yet and does not know the answer.

Write ONE paragraph of natural first-person reasoning (70-220 words, hard cap 300) that:
- starts from what the question requires (the needs), not from a frame-by-frame description;
- states what the overview establishes, citing the timestamps that matter (use "at 105.0 s" style, only timestamps from the material), and distinguishes what is directly visible from what is inferred across frames;
- names the gap precisely: which need is still open, what exactly is not readable/visible, and why that leaves several answers possible — without naming any option letter. Never state a value for the attribute an open need is about, not even tentatively ("appears to read DBS", "looks white"): describe only what is resolvable at overview scale ("two logos, letters too small to read"). A first turn that already contains the decisive fact makes the crop decorative;
- for a crop: derives each window from what the frames show — an anchor frame that is the moment, or a candidate frame that merely could be it — (why it starts where it starts and ends where it ends) and says what a close look at that range should reveal; with two or three windows, present them as competing hypotheses: what makes each a plausible location of the reference event, and that the close look will decide which one is the moment — never assert in advance which is right;
- for a direct answer (route DIRECT): argues from the cited frames to the unique answer by content — the identity of the moment, the attribute read at the needed precision, the fact that rules out the closest alternative — names the answer content in words (never the letter), and ends by stating that the overview already settles every need. No gap or hedge vocabulary ("cannot tell", "unclear", "not visible", "closer look") belongs in a direct turn, not even negated.
The range(s) you derive must be exactly the window(s) given in the material (start and end in seconds) — the agent calls those numbers right after this paragraph.
Forbidden: option letters; facts not in the material; "the tool"/"tool call"/"the model" meta talk; filler ("let me", "I will now", "clearly", "comprehensively"); hedging theatre. The reasoning must be the kind an expert writes when thinking, dense and specific.
Output the paragraph only.

Figure E.10: **Prompt for Assemble Stage of Data Curation**

F More Cases

To better visualize the behavior of WORTH, we present its actual outputs in Figure F.11 to Figure F.14. Each episode shows the overview that the policy receives, the question with the correct option marked in red, the requests that the policy issues, the frames of every requested window as inspected by the reader, the returned observations, and the final answer. The episodes cover the four interaction

patterns of the interface, namely answering directly from the overview, a single request, several requests issued in parallel within one plan, and two plans in one episode



Figure F.11: **Answering directly from the overview.**

Case 1: 32 uniformly sampled frames (one per 1s in this 32s video.)



Which of the following objects the camera wearer of the video did NOT interact with?
A. fridge door. B. food processing machine. C. fridge. D. lid of bottle. E. squash

<tool_call>{"task": "Look carefully at all kitchen appliances on the counters and island. Is there a food processor or blender visible? Does the person touch, open, or operate any such machine? Also check if the person handles or cuts the squash/fruit on the cutting board.", "start_time": 10.0, "end_time": 25.0}</tool_call>



<tool_response>Inspected 10.0-25.0 s.
- 10.0-12.3 s: Person stands at the kitchen island/counter pouring a pink/red liquid from a plastic bottle (blue cap) into a clear glass; a light-blue/white cutting board sits on the right side of the counter with a small green item (a piece of squash/fruit) resting on it.
- 13.5 s: Person opens a refrigerator (door swung open, shelves with jars/eggs visible) and places the bottle inside; no blender/food processor involved.
- 14.6 s: Wider kitchen view shows the microwave, a paper-towel roll, and a dish rack holding plates and pots on the left counter; still no food processor or blender anywhere.
- 18.1-23.8 s: Person moves to the sink and fills the glass with water from the tap; the cutting board with the green item is out of view and untouched.
- 25.0 s: Person lifts the water-filled glass to drink; dish rack and counter visible, no blender/food processor
</tool_response>

<answer>B</answer> ✓ B. food processing machine.

Case 2: 64 uniformly sampled frames (one per 12.2s in this 771s video.)



What animal does "Xuanxuan" play?
A. Turtle. B. Dinosaur. C. Alligator. D. Frog. E. Crab. F. Snail. G. Lizard.

<tool_call>{"task": "Watch the child labeled 轩轩 (Xuanxuan) in the yellow puffy costume with green limbs. Describe the costume precisely: shape of the back (flat shell vs spiky), number/position of limbs, head shape, and any on-screen text, caption, or prop that names the animal he is playing. Note whether it reads as a turtle, dinosaur, or another creature. ", "start_time": 110.0, "end_time": 150.0}</tool_call>



<tool_response>Inspected 110.0-150.0 s.
- 119.2-128.4 s: A separate adult in a full green turtle costume (green face mask, turtle shell, beaded necklace) appears — this is the 'village chief'/dad figure, distinct from the child 轩轩.
- 121.3-147.9 s: The child's real head is exposed (wet dark hair, no animal headpiece or helmet), so the 'head' of the creature is just the child's own head.
- 129.5-130.5 s: The child in the yellow puffy costume is identified by on-screen caption '轩轩' (Xuanxuan) appearing at lower-left next to him.
- 133.6-137.7 s: On-screen caption names the animal: '大竣 这忍者神龟 听着像你爸爸' (Da Jun: this ninja turtle sounds like your dad) — i.e. the costume is a 忍者神龟 / Ninja Turtle.
- 143.8-145.9 s: Reinforcing caption '那个忍者神龟' (that ninja turtle) appears while the yellow-costumed child is shown.
- 146.9-148.9 s: Caption '早已看穿一切的样子 / 其实不是我爸爸 是"村长"' (looks like he's seen through everything; actually not my dad, it's the 'village chief') confirms the turtle role is being played in a skit.
</tool_response>

<answer>A</answer> ✓ A. Turtle.

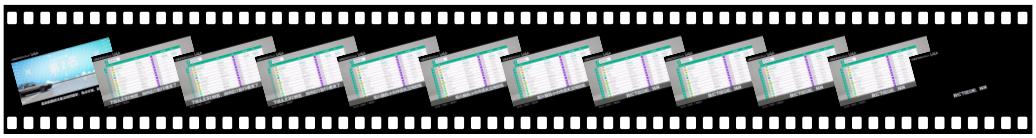
Figure F.12: **A single request.**

Case 1: 64 uniformly sampled frames (one per 10.3s in this 648.5s video.)



What color is the car that ultimately takes first place in the video?
Options:
A. Black. B. White. C. Blue. D. Red. E. Yellow. F. Silver. G. Orange. H. Green.

<tool_call>{"task": "Read the final standings/leaderboard table: identify the car in the #1 row and report its body color, and note the car name/icon. Confirm whether the top car matches the white player car.", "start_time": 630.0, "end_time": 649.0}</tool_call>



<tool_response>Inspected 630.0-649.0 s.
- 630.0-631.1 s: A white two-door hatchback with a black lower body (Toyota Trueno/AE86 livery) is shown on a checkered start line in front of a blue '第2名' (2nd place) banner — this is the player's car, which finished 2nd.
- 632.2-648.5 s: A final standings table titled '歌利亚' (Goliath) is displayed with columns 名次(rank), 车手(player), 车辆(vehicle), 等级(grade), 最佳单圈(best lap), 时间(time).\\n- 632.2-648.5 s: Row #1 (01): player 'TheMonsterMash' with orange icon '22', vehicle '日产 370Z' (Nissan 370Z), grade S1 897, best lap 10:20.494, time 10:23.582.
- 632.2-648.5 s: Row #2 (02): player 'Mr.Lootbox' (highlighted, with a green dot), vehicle '丰田 Trueno' (Toyota Trueno), grade S1 900, best lap 09:59.356, time 10:23.923 — this matches the white player car.
- 643.0-648.5 s: The 370Z (the #1 car) is never rendered as a 3D model in this segment; only the white Trueno player car appears on screen, so the 370Z's body color cannot be observed.
- 648.5 s: Closing subtitle '我们下回见啦，拜拜' (see you next time, bye) appears as the table fades to black
</tool_response>

<tool_call>{"task": "Identify the Nissan 370Z among the cars on the road at the start of the race and report its body color. The 370Z is the #1 finisher; it should be visible as a distinct car (not the white Trueno) during the early race segment.", "start_time": 0.0, "end_time": 120.0}</tool_call>



<tool_response>Inspected 0.0-120.0 s; closer look at 0.0-12.0 s.
- 0.0-1.1 s: Player car is a white Toyota Trueno (AE86) with yellow plate reading 'RED SUNS', viewed from behind on a two-lane road beside a lake with a stone wall on the left.
- 0.0-1.1 s: Ahead in the right lane is a red muscle car (Camaro-style, wide/boxy rear) with a yellow plate.
- 0.0-1.1 s: Ahead in the center/left lane further down is a second red car.
- 0.0-1.1 s: Ahead in the center distance is a blue sports coupe; this is the Nissan 370Z (the #1 finisher). Its body color is blue.
- 0.0-1.1 s: A row of additional start-line cars (red, blue, etc.) is visible far ahead near the start gantry.
- 2.2-12.0 s: On-screen text '让它们20秒' (let them 20 seconds) appears; HUD shows lap 0/1, position 12/12, and a rising lap timer, confirming the race start.
- 12.0-120.0 s: The white Trueno accelerates away from the start, passing a festival gantry, checkpoints, and village roads; the blue 370Z is no longer visible ahead after the opening seconds.
</tool_response>

<answer>C</answer> ✓ C. Blue.

Figure F.13: **Two plans in one episode.**

Case 1: 64 uniformly sampled frames (one per 4.76s in this 300.0s video.)



The following 2 people appeared in the video: 1. The individual is a male wearing a light-colored trench coat and khaki pants. He has short hair and glasses. His shoes are black, and he appears to be holding something in his hand, possibly a phone or similar object. 2. The individual is male, wearing a bright orange jacket and dark blue jeans. He has short dark hair and is wearing black shoes. His posture is upright as he walks forward. From which edge of the frame did they enter the scene
A. 1 entered through the top edge, 2 entered through the left edge
B. 1 entered through the top edge, 2 entered through the right edge
C. 1 entered through the right edge, 2 entered through the top edge
D. both entered through the right edge\\nE. both entered through the left edge

<tool_call>{"task": "Track the man in the light/cream trench coat and khaki pants (short hair, glasses). Identify the frame edge he first appears from (top/background, left, or right) and his direction of travel at that moment.", "start_time": 80.0, "end_time": 100.0}</tool_call>
<tool_call>{"task": "Track the man in the bright orange jacket and dark blue jeans. Identify the frame edge he first appears from (top/background, left, or right) and his direction of travel at that moment.", "start_time": 280.0, "end_time": 300.0}</tool_call>



<tool_response>Inspected 80.0-100.0 s. - 80.0-91.6 s: A man in a light/cream trench coat and khaki/tan pants (short hair, glasses) is standing still on the wide stone steps of the building in the upper-right background; he does not move during this span. - 80.0-100.0 s: He is present from the very first frame of this segment, positioned in the top/background (upper-right on the steps); he does not enter from a side edge during this segment. - 92.6-95.8 s: The same man begins descending the steps, moving from the upper-right toward the lower-left, reaching the bottom of the steps by ~95.8s. - 96.8-100.0 s: He is now on the open plaza, walking leftward (toward the left side of the frame), passing in front of a woman in a dark coat who is standing in the middle of the plaza.</tool_response>
<tool_response>Inspected 280.0-300.0 s. - 280.0-281.2 s: At 280.0s the man in the bright orange jacket and dark blue jeans is first visible at the right edge of the frame (right side, roughly mid-height), already in motion. - 280.0-282.4 s: At his first appearance he is moving leftward and away from the camera (toward the background), i.e. entering from the right edge and traveling toward the center/background. - 280.0-300.0 s: Throughout the segment his direction of travel is consistently away from the camera (into the background), with a leftward drift; he never moves toward the camera. - 282.4-285.9 s: He continues walking away from the camera toward the background, now in the center of the frame, accompanied by a man in a black coat and light jeans to his right. - 287.1-300.0 s: He keeps walking away from the camera and drifts toward the left side of the frame, becoming smaller in the distance as the segment ends. </tool_response>

<answer>B</answer> ✓ B. 1 entered through the top edge, 2 entered through the right

Figure F.14: **Two parallel requests.**