FROM SCAN TO FOCUS: LEARNING A VALUE-GUIDED ADAPTIVE BROWSING POLICY FOR EFFICIENT LONG-VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video reasoning requires Multimodal Large Language Models (MLLMs) to locate sparse evidence over extended temporal contexts while preserving the fine-grained visual details needed for reliable answers. Uniform sparse sampling can miss brief but decisive events, whereas dense processing incurs substantial visual-token and memory costs. To address this challenge, we introduce MindSight, a human-inspired adaptive browsing agent that treats long-video understanding as active perception. MindSight follows a scan–focus–reflect process: it first builds a low-cost global view to identify candidate intervals; conditioned on a candidate interval and the evidence acquired so far, it then compares STOP, LOW, and HIGH by their expected downstream returns, deciding whether to stop or inspect the interval at low or high visual fidelity by trading off new evidence against its observation cost. We learn this observation-value decision through paired Monte Carlo continuations at matched timestamps, and optimize multi-turn browsing trajectories with turn-balanced geometric-mean policy optimization (TB-GMPO). Experiments with two backbones on four long-video benchmarks improve accuracy by 1.2–12.0 percentage points (pp) over the corresponding base models, while reducing visual-token consumption and improving inference efficiency.

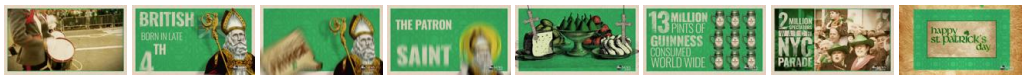
1 INTRODUCTION

Long video reasoning poses a challenge for Multimodal Large Language Models (MLLMs): how to efficiently acquire and reason over relevant evidence from extended temporal contexts, where crucial information is often temporally sparse yet visually detailed. Unlike short videos, long videos may contain decisive events lasting only seconds within minutes or hours of irrelevant content. Existing MLLMs face a critical trade-off: sparse frame sampling may overlook fleeting but essential evidence due to inaccurate temporal grounding, whereas dense processing of the entire video introduces substantial visual-token and memory overhead. Therefore, successful reasoning requires not only recognizing visual content, but also determining *where to look*, *how much visual detail to acquire*, and *when sufficient evidence has been collected*.

Existing approaches reduce the cost of long-video reasoning through longer context modeling, visual-token compression, frame retrieval, and memory, but fixed representations can miss subtle cues or evidence that becomes relevant only during reasoning (Maaz et al., 2024; Lin et al., 2024a; Song et al., 2024; Zhang et al., 2024b). Recent video agents address *where to look* through iterative temporal navigation, evidence retrieval, or adaptive visual acquisition (Wang et al., 2024b; Yuan et al., 2025; Qiu et al., 2026; Ge et al., 2025; Wang et al., 2026c; Ding et al., 2025). Among them, VideoZoomer jointly selects a temporal interval and frame rate (Ding et al., 2025). These advances establish adaptive video exploration, but leave open how to explicitly supervise the marginal value of stopping versus acquiring low- or high-resolution evidence at the same candidate interval.

Interestingly, humans naturally address this challenge through adaptive browsing. As illustrated in Figure 1, to answer *"What happened when Irish immigrants brought St. Patrick's Day to America?"*, a viewer can: (1) perform a **Global Scan** at low cost to build context and identify candidate moments; (2) apply **Value-Guided Focus** to a selected segment at the detail needed to resolve the current uncertainty; and (3) **Reflect** on the new evidence to decide whether it is sufficient or fur-

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.



0s 30s 60s 90s 120s 150s 180s 210s

Question and answer options

What happened when Irish immigrants brought the tradition of St. Patrick's Day to America?
A. Centered around drinking and celebrating; B. St. Patrick moved to America; C. Celebrated St. Patrick's birth instead; D. Attended church and gathered for feasts.

1. Global Scan (LOW, at most 64 frames, full video)

<action> SCAN </action>
<caption> Historical overview of St. Patrick's Day in Ireland and America. </caption>
<timeline> marching band [0–30s]; history [30–60s]; immigrants and U.S. traditions [60–120s]; parades and statistics [120–180s]; closing [180–210s]. </timeline>
<candidates> [35–45s]: historical context; [60–90s]: immigrants and U.S. traditions. </candidates>
<think> A:8, B:2, C:1, D:3. The American transition remains unresolved; [60–90s] is the strongest candidate. </think>
<reason> Prioritize evidence that distinguishes American celebrations from religious alternatives. </reason>

2. Value-Guided Focus

<candidate> [60–90s], sampled with the 24-frame cap </candidate>
<action> HIGH on [60–90s] </action>
<reason> Scene text can resolve the alternatives, so its expected gain justifies HIGH's cost. </reason>
<observation> Immigrants brought the tradition to America, where it became associated with public celebration and drinking. </observation>

3. Reflect: Update Evidence Sufficiency

<evidence> Focused text supports A; B–D are not supported by the observed account. </evidence>
<option-support> A:0.8, B:0.1, C:0.2, D:0.3 </option-support>
<action> STOP </action>
<reason> Evidence is sufficient for A; another observation is not worth its cost. </reason>

4. Stop and Answer

<answer> A </answer>
<reason> Focused evidence resolves the historical transition. </reason>

Figure 1: **Motivating scan–focus–reflect example.** Scan proposes candidates, value-guided focus selects HIGH, and reflection stops when further observation is not worth its cost.

ther observation is worthwhile. This *scan–focus–reflect* process directs visual computation toward evidence that can resolve the remaining uncertainty.

Inspired by this human browsing strategy, we propose **MindSight**, a human-inspired adaptive browsing agent that reformulates long-video reasoning as active perception. Conditioned on a candidate interval and current evidence, MindSight compares STOP, LOW, and HIGH by their expected downstream returns, jointly deciding *where to look, whether finer visual detail is worth its cost, and when to stop*. As standard video question–answer supervision lacks browsing trajectories, we bootstrap and replay-verify multi-step trajectories that link observation requests to their resulting evidence. Paired Monte Carlo continuations at matched timestamps estimate the long-term value of each action, while SFT establishes the browsing protocol and TB-GMPO optimizes the complete multi-turn trajectories. Our main contributions can be summarized as follows:

- We propose **MindSight**, a human-inspired adaptive browsing agent that jointly decides *where to look, whether acquiring finer visual detail is worth its cost, and when to stop*. By separating temporal proposal from observation value, MindSight allocates resolution according to the current evidence and the expected long-term return of each action.
- We introduce paired Monte Carlo continuation supervision to estimate the downstream value of STOP, LOW, and HIGH at matched timestamps. We also propose TB-GMPO to

optimize multi-turn trajectories with balanced turn contributions and guarded likelihood-ratio updates.

- Experiments with two backbones on four long-video benchmarks show consistent gains of 1.2–12.0 pp over the corresponding base models. In our controlled A100 serving setting, MindSight also achieves lower TTFT and peak memory than uniform dense inference.

2 Related Work

Long-video understanding with MLLMs. Early video-language systems emphasized instruction tuning and temporal representation (Maaz et al., 2024; Lin et al., 2024a; Song et al., 2024), while recent MLLMs improve long-context processing through large-scale video instruction data, efficient visual encoders, token compression, and long-context architectures (Zhang et al., 2025b; Lin et al., 2024b; Liu et al., 2025; Zhang et al., 2024b; Shen et al., 2025; Wang et al., 2025c; Ryoo et al., 2025; Zhang et al., 2025a; Cheng et al., 2025; Wang et al., 2024a; Li et al., 2025a). Another line of work retrieves relevant frames, grounds temporal evidence, or compresses videos into textual and hierarchical memories before answering (Wang et al., 2026a; Yan et al., 2026b; Liu et al., 2026; Li et al., 2024; Ren et al., 2024; Zhang et al., 2024a; Wang et al., 2024c; 2025d). However, fixed representations and one-shot retrieval can miss fine-grained or newly relevant evidence. MindSight instead revisits the video sequentially at an adaptive granularity.

Agentic video reasoning and policy learning. Video agents extend long-video understanding with iterative planning, retrieval, tool use, verification, and memory (Wang et al., 2024b; Fan et al., 2024; Kugo et al., 2024; Chen et al., 2026a). Recent systems learn temporal navigation, multi-step clue seeking, adaptive zooming, and evidence-guided reasoning (Qiu et al., 2026; Yang et al., 2026; Yin et al., 2026; Ding et al., 2025; Nie et al., 2026; Ge et al., 2025; Wang et al., 2026c). Hypothesis-verification and rubric-guided approaches make discriminative or missing evidence explicit (Wang et al., 2026b; Yan et al., 2026a), and budget-aware agents learn when to continue observing or terminate (Liang et al., 2026). Reinforcement learning further improves video reasoning through answer, format, temporal-grounding, and tool-use rewards (Li et al., 2025b; Wang & Peng, 2025; Dang et al., 2025; Wang et al., 2025b; Feng et al., 2025; Li et al., 2025c; Chen et al., 2026b; 2025a). VideoZoomer learns where to revisit and at what frame rate (Ding et al., 2025). MindSight instead uses paired Monte Carlo continuations to estimate cost-aware returns for STOP, LOW, and HIGH at each state and proposed interval, while TB-GMPO optimizes the multi-turn policy.

3 Methodology

3.1 Overview: Browsing by the Value of Additional Evidence

A fixed low-resolution scan initializes the question-conditioned state h_t. We factor each browsing decision into temporal proposal and observation-value selection:

$$\pi_\theta(s_t, a_t \mid h_t) = \pi_{\text{loc}}(s_t \mid h_t)\, \pi_{\text{value}}(a_t \mid h_t, s_t), \qquad a_t \in \{\text{STOP}, \text{LOW}, \text{HIGH}\}, \tag{1}$$

where s_t proposes where to inspect and a_t decides whether the expected evidence warrants no observation (STOP), low resolution (LOW), or high resolution (HIGH). Both factors share one policy; the factorization describes their decision order rather than statistical independence. LOW and HIGH use identical timestamps at 224^2 and 448^2, isolating observation fidelity from temporal localization. Each observation updates the evidence state before the next decision. Appendix A gives the execution protocol.

3.2 Stage 1: Browsing Policy Learning

3.2.1 Trajectory Bootstrapping.

We construct the SFT pool from LongVideo-Reason (Chen et al., 2025a), which contains 52K high-quality long-video question–reasoning–answer pairs. We randomly sample 25K examples from its longer-video portion and generate browsing trajectories for them. For each video–question pair, the explorer samples a trajectory $\tau = \{(s_t, a_t, o_t)\}_{t=1}^{T}$ and final answer $\hat{y}$. We discard trajectories with

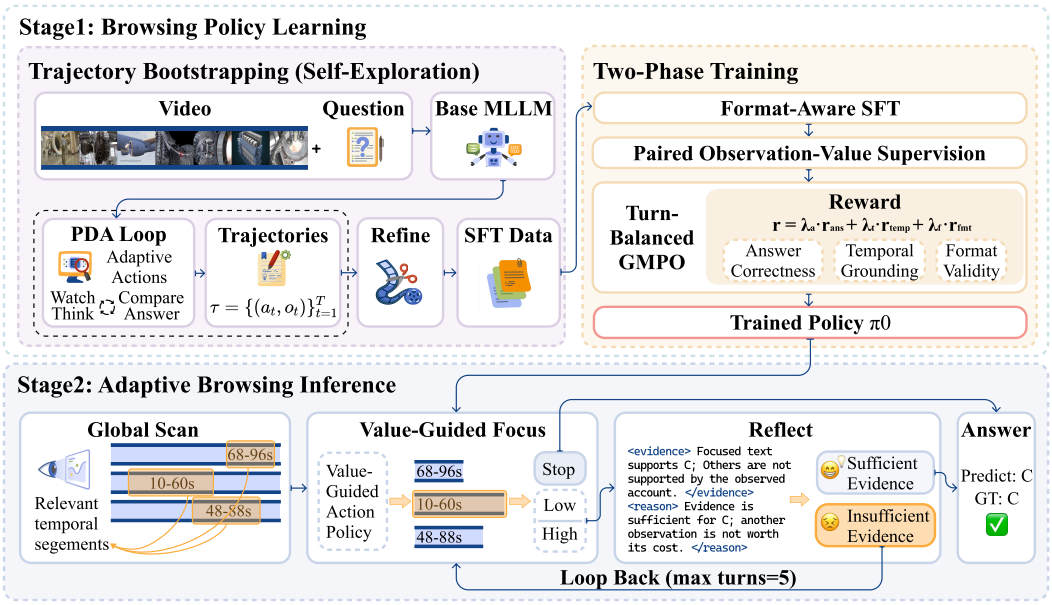


Figure 2: **Overview of MindSight.** Stage 1 bootstraps replay-verified browsing trajectories through self-exploration. Format-aware SFT initializes the action protocol, paired observation-value supervision teaches the policy to compare STOP, LOW, and HIGH at matched candidate intervals, and TB-GMPO optimizes complete trajectories using answer, temporal-grounding, and format rewards. During Stage 2, a global scan proposes relevant intervals, value-guided focus decides whether and at what fidelity to inspect, and reflection either answers with sufficient evidence or updates the state and loops back for another observation.

incorrect final answers and retain a short executable subsequence τ^* from each successful trajectory, yielding $\mathcal{D}_{\text{SFT}} = \{(V, Q, \tau^*, y^*)\}$. We preserve action–observation dependencies and replay every retained trajectory for verification, resulting in approximately 3K SFT trajectories.

3.2.2 Format-Aware Supervised Fine-Tuning.

Each retained trajectory is serialized as observation–response pairs (u_j, r_j). SFT minimizes

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\tau^* \sim \mathcal{D}_{\text{SFT}}} \sum_j \log p_\theta(r_j \mid u_{\le j}, r_{<j}), \tag{2}$$

over assistant-generated tokens only, initializing the action protocol while treating observations as context.

3.2.3 Paired Observation-Value Supervision.

We first train an initial SFT policy on replay-verified browsing trajectories and freeze a snapshot to generate counterfactual continuations. At a training state h with proposed interval s, the snapshot rolls out STOP, LOW, and HIGH under the same remaining budget. LOW and HIGH inspect the same timestamps and differ only in resolution. We score each branch by its final-answer correctness and incremental observation cost:

$$V(h, s, a) = \mathbb{E}[\mathbb{I}(\hat{y}_a = y^*) \mid h, s, a] - \lambda_C \mathbb{E}[C_a \mid h, s, a], \tag{3}$$

For continuation m, we define the normalized branch cost as

$$C_{a,m} = \frac{N_{\text{acq}}(\tau_{a,m}) - N_{\text{acq}}(h)}{B_{\text{vis}}}, \tag{4}$$

where $N_{\text{acq}}(h)$ is the visual-token cost already incurred at the branch state and $N_{\text{acq}}(\tau_{a,m})$ is the cumulative acquired visual-token cost when that continuation terminates. The numerator therefore

includes every new observation from the forced action onward, including later continuation actions and repeated observations. The shared scan is excluded because it is identical across branches. We use the fixed maximum focus budget $B_{\text{vis}} = KN_{\text{HIGH}} = 5 \times 3{,}072 = 15{,}360$ tokens. STOP has zero incremental observation cost. A failed request that returns no visual tokens also adds zero visual cost but still consumes one browsing attempt.

With M continuations, we estimate $\hat{V}_a = M^{-1} \sum_m [z_{a,m} - \lambda_C C_{a,m}]$, where $z_{a,m} = \mathbb{I}(\hat{y}_{a,m} = y^*)$. A state contributes a hard target a^* only if its lower confidence bound exceeds every alternative's upper bound by τ_V; ambiguous states are omitted. Continued SFT combines the original trajectory loss for interval proposal with action-token cross-entropy for STOP/LOW/HIGH, and the resulting checkpoint initializes TB-GMPO.

Matched LOW/HIGH timestamps isolate fidelity from temporal localization, so the target asks whether additional evidence justifies its cost. All training branches share M, decoding, remaining budget, and answer extraction, and use training-set answers only. Counterfactual branching is used for training targets and offline evaluation diagnostics only; deployed inference executes one action. We select the cost unit, λ_C, M, and τ_V on the development set.

3.2.4 Turn-Balanced Geometric-Mean Policy Optimization.

Long trajectories mix brief actions with verbose reasoning, so token averaging can overweight long reflections and extreme likelihood ratios can dominate updates. We adapt geometric-mean policy optimization (Zhao et al., 2026) with equal turn weights and bounded log ratios, yielding *turn-balanced geometric-mean policy optimization* (TB-GMPO).

Trajectory return. For a group of G complete trajectories sampled from π_{old}, the terminal Monte Carlo return is

$$R_i = \lambda_a R_{\text{ans},i} + \lambda_t R_{\text{temp},i} + \lambda_f R_{\text{fmt},i}. \tag{5}$$

Here $R_{\text{ans}} = \mathbb{I}(\hat{y} = y^*)$, and the format reward averages XML validity over the M_i assistant turns,

$$R_{\text{fmt},i} = \frac{1}{M_i} \sum_{j=1}^{M_i} \mathbb{I}(\text{ValidXML}_{ij}). \tag{6}$$

The temporal reward is $R_{\text{temp}} = \text{IoU}(\hat{S}, S^*) - \gamma|\hat{S}|/|V|$ when temporal labels are available. The set $\hat{S}$ contains successful focus intervals but excludes the global scan. This return supervises complete trajectories; Equation 3 separately supervises STOP/LOW/HIGH. We use $(\lambda_a, \lambda_t, \lambda_f) = (0.5, 0.2, 0.3)$ and $\gamma = 0.01$.

Action-aligned aggregation. Let L_{ij} be the number of generated assistant tokens in turn j of trajectory i, and M_i its number of nonempty assistant turns. A turn is defined by one runtime-validated executable action plus its preceding assessment; model-emitted delimiters cannot create extra turns. User, tool, visual, and padding tokens are masked out. Define

$$d_{ijt} = \log \pi_\theta(x_{ijt} \mid h_{ijt}) - \log \pi_{\text{old}}(x_{ijt} \mid h_{ijt}), \qquad w_{ijt} = \frac{1}{M_i L_{ij}}. \tag{7}$$

Thus $\sum_{j,t} w_{ijt} = 1$ and every turn receives weight $1/M_i$, regardless of length. We compute the detached advantage

$$A_i = \text{clip}\left(\frac{R_i - \mu_R}{\sigma_R + \delta}, -A_{\max}, A_{\max} \right), \quad \delta > 0, \tag{8}$$

where μ_R and σ_R are the within-group mean and population standard deviation. We use $\delta = 10^{-6}$ and $A_{\max} = 2$. We guard ratio outliers with $u_{ijt} = c \tanh(d_{ijt}/c)$, where $c = 1.5$, and apply GMPO-style token-level pessimistic clipping before aggregation:

$$v_{ijt} = \begin{cases} \min(u_{ijt}, \log(1+\epsilon)), & A_i > 0, \\ \max(u_{ijt}, \log(1-\epsilon)), & A_i < 0, \end{cases} \qquad g_i = \exp\left(\sum_{j,t} w_{ijt} v_{ijt} \right). \tag{9}$$

The parameter-gradient norm is clipped at 10 in all experiments. For $A_i = 0$ the policy contribution is zero. The objective is

$$\mathcal{J}_{\mathrm{TB}} = \mathbb{E}\left[\frac{1}{G}\sum_i A_i g_i\right] - \beta\mathcal{K}(\pi_\theta, \pi_{\mathrm{ref}}), \tag{10}$$

where $\mathcal{K}$ is the turn-weighted conditional KL to the frozen SFT policy. Token-level clipping preserves gradients from unaffected tokens. With uniform token weights and $c \to \infty$, the policy term recovers the GMPO surrogate under the same clipping thresholds.

What stability is guaranteed. With fixed detached weights and advantage, the policy term $J_i = A_i g_i$ satisfies, away from clipping boundaries,

$$\frac{\partial J_i}{\partial d_{ijt}} = A_i g_i w_{ijt} m_{ijt} \operatorname{sech}^2(d_{ijt}/c), \qquad \sum_{j,t}\left|\frac{\partial J_i}{\partial d_{ijt}}\right| \le A_{\max} e^c, \tag{11}$$

where $m_{ijt} \in \{0, 1\}$ indicates an active clipping branch. The bound is independent of response length but does not guarantee lower variance, convergence, or accuracy; the guard can also attenuate useful corrections. Appendix B provides the branch-wise analysis.

3.3 Stage 2: Adaptive Browsing Inference

Scan and value-guided action. Given video V and question Q, MindSight applies the same low-resolution scan used in training, sampling at most 64 frames to initialize a timestamped timeline, provisional answer, evidence, unresolved claims, and remaining budget. At turn t, the shared policy proposes interval s_t and selects $a_t \in \{\mathrm{STOP}, \mathrm{LOW}, \mathrm{HIGH}\}$ from (h_t, s_t). STOP answers without new frames; LOW and HIGH inspect identical timestamps at 224^2 and 448^2. Paired targets teach whether another observation is worth its cost and whether HIGH fidelity is needed. Figure 3 compares this policy with fixed LOW/HIGH and confidence stopping.

Closed-loop update and consistency. For LOW or HIGH, the runtime validates and executes the request, then appends returned frames and timestamps in the training-time dialogue format. MindSight updates its provisional answer, evidence, and unresolved claims before the next proposal, so later actions depend on actual observations. The loop ends at STOP, the shared limit $K = 5$, or a runtime limit. Training and inference share processors, frame caps, action grammar, validation, serialization, and state transitions. Deployed inference never branches; offline evaluation may run matched counterfactuals only for diagnostics. Algorithm 1 gives the complete protocol.

4 Experiments

4.1 Experimental Setup

Benchmarks and training data. We report multiple-choice accuracy on MLVU (Zhou et al., 2025), LVBench (Wang et al., 2025a), VideoMME (Fu et al., 2025), and LongVideoBench (Wu et al., 2024). We use the 30–60-minute VideoMME-Long subset and the 15–60-minute LongVideoBench tier, following official aggregation. From LongVideo-Reason's 52K pairs (Chen et al., 2025a), we sample 25K longer-video examples for trajectory generation; filtering and replay verification yield about 3K SFT trajectories, and paired branching uses only this SFT pool. We separately sample 10K non-overlapping video–question examples for TB-GMPO and split them once 80/20 for training and held-out validation, shared across optimizer variants. Figure 4 reports accuracy on the validation split; no example is shared between the SFT and RL pools.

Models and protocol. The backbones are Qwen2.5-VL-7B-Instruct (Bai et al., 2025a) and Qwen3-VL-8B-Instruct (Bai et al., 2025b). All reported MindSight results are means over three independent training seeds (42, 100, and 250) and use the current pipeline: replay-verified SFT with paired action supervision followed by TB-GMPO. Scan/focus budgets and processor-aware costs are defined in Section 4.5.

Implementation. SFT runs for one epoch at learning rate 10^{-5}. TB-GMPO runs for five epochs (10,000 optimizer updates) with $G = 8$, learning rate 5×10^{-6}, $\epsilon = 0.2$, $\beta = 0.1$, ratio guard

Table 1: **Main results on long video reasoning benchmarks.** MindSight results are means over three independent training seeds (42, 100, and 250). Params denotes language-model size. VideoZoomer provides a parameter-matched 7B agent comparison; its unmarked results are reported by the original paper under its inference budget. Results marked with † are obtained from released checkpoints under our unified protocol. Bold and underlining indicate the highest and second-highest listed open-source values per column.

Model	Params	MLVU (Test, %)	LVBench (%)	VideoMME-Long (%)	LongVideoBench (%)
Closed-source Models					
GPT-4o (OpenAI, 2024)	–	54.9	48.9	72.1	66.7
Gemini-1.5-Pro (Reid et al., 2024)	–	–	33.1	77.4	64.0
Seed1.5-VL-Pro (Guo et al., 2025)	–	54.9	46.1	63.3	63.7
Open-source Vanilla Models					
LongVILA (Chen et al., 2025b)	7B	49.0	–	52.1	57.7
Video-XL (Shu et al., 2025)	7B	45.5	–	54.9	50.7
InternVL-2.5 (Chen et al., 2025c)	8B	<u>68.9</u>	23.9	53.1	60.0
InternVL-3 (Zhu et al., 2025)	8B	**71.4**	25.9	–	58.8
Qwen2.5-VL-Instruct (Bai et al., 2025a)	7B	48.0	38.3	50.0	56.7
Qwen3-VL-Instruct (Bai et al., 2025b)	8B	53.5	40.5	54.0	58.5
Open-source Reasoning Models					
Temporal-R1 (Li et al., 2025b)	7B	56.4†	32.7	–	–
Open-R1-Video (Wang & Peng, 2025)	7B	–	26.9	50.8†	–
TW-GRPO (Dang et al., 2025)	7B	55.6†	30.0	51.5†	–
Video-R1 (Feng et al., 2025)	7B	–	34.7	52.2†	52.0†
Time-R1 (Wang et al., 2025b)	7B	56.9†	32.7	53.1†	–
VideoChat-R1 (Li et al., 2025c)	7B	55.3†	33.0	–	55.1†
GRPO-CARE (Chen et al., 2026b)	7B	–	33.3	50.5†	53.7†
Open-source Video Agents					
VideoExplorer (Yuan et al., 2025)	7B	58.6	33.3	–	–
VideoZoomer (Ding et al., 2025)	7B	55.8	<u>41.5</u>	55.8	57.7
MindSight (Ours)	7B	59.5	39.5	56.0	<u>60.5</u>
MindSight (Ours)	8B	65.5	**43.5**	**59.0**	**63.5**

$c = 1.5$, advantage cap $A_{\max} = 2$, and reward weights $(\lambda_a, \lambda_t, \lambda_f) = (0.5, 0.2, 0.3)$. RL rollouts use temperature 1.0 and top-$p = 0.95$; evaluation uses vLLM with tensor parallelism 4, temperature 0.1, and top-$p = 0.001$. Training and inference share a 64K context and at most five browsing attempts. Table 4 summarizes additional implementation details, and Appendix C gives the measurement protocol.

4.2 Main Results

Table 1 shows that MindSight improves its corresponding backbone on every benchmark: the 7B and 8B variants gain 1.2–11.5 pp and 3.0–12.0 pp, respectively. Among the reported numbers, MindSight-7B exceeds the parameter-matched VideoZoomer on MLVU, VideoMME-Long, and LongVideoBench, while VideoZoomer leads on LVBench. Scaling MindSight to 8B yields the best listed open-source results on LVBench, VideoMME-Long, and LongVideoBench. The VideoZoomer comparison matches parameter scale but not inference budget, so it should be interpreted as a published-result comparison rather than a controlled efficiency comparison.

4.3 Ablation Study

Table 2 shows stable three-seed performance: MindSight reaches 39.5% on LVBench and 56.0% on VideoMME-Long, both with a 0.2 pp standard deviation. Removing paired value supervision lowers accuracy by 1.1/1.8 pp, while replacing TB-GMPO with GMPO lowers it by 2.8/3.7 pp. GMPO and variants without turn balancing or the log-ratio guard also show higher variation (0.4–0.6 vs. 0.2 pp). Figure 4 further shows controlled ratio tails, lower pre-clipping gradient norms, and less KL growth. Single-pass and uniform-resolution ablations confirm the value of iterative, multi-resolution inference.

Compared with confidence stopping, paired value supervision improves accuracy by 0.7/0.9 pp on LVBench/VideoMME-Long, uses 43.1/23.6% fewer visual tokens, and reduces false STOP from

Table 2: **Component analysis on LVBench and VideoMME-Long.** Entries report mean accuracy (%) ± standard deviation (pp) over three independent training seeds (42, 100, 250). All variants use Qwen2.5-VL-7B-Instruct and the same evaluation protocol.

Model Variant	LVBench (%)	VideoMME-Long (%)
Qwen2.5-VL-7B-Instruct (Base)	38.3 ± 0.1	50.0 ± 0.1
Training components		
w/o SFT	36.1 ± 0.3	51.7 ± 0.3
w/o policy optimization	34.8 ± 0.3	51.1 ± 0.3
w/o paired value supervision	38.4 ± 0.2	54.2 ± 0.2
w/o temporal reward (λ_t=0)	38.8 ± 0.2	55.0 ± 0.2
Policy optimization variants		
GMPO (replace TB-GMPO)	36.7 ± 0.5	52.3 ± 0.5
w/o turn-balanced weighting	36.8 ± 0.4	52.9 ± 0.4
w/o log-ratio guard	37.0 ± 0.6	53.3 ± 0.6
Inference components		
w/o multi-turn browsing (single-pass)	33.8 ± 0.2	49.7 ± 0.2
w/o multi-resolution (uniform res.)	37.0 ± 0.2	53.6 ± 0.2
MindSight (Full)	$\mathbf{39.5 \pm 0.2}$	$\mathbf{56.0 \pm 0.2}$

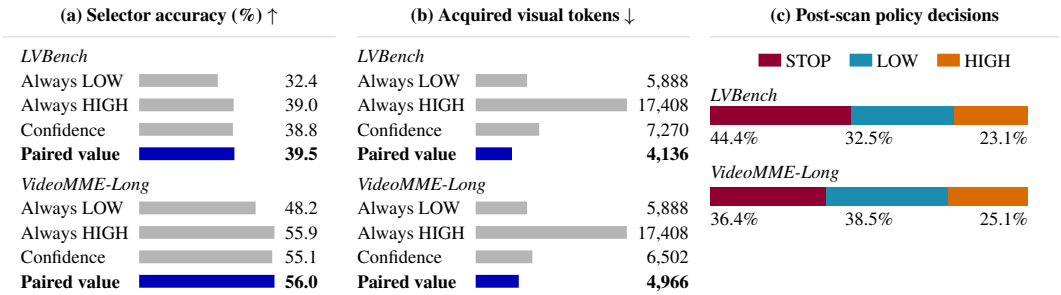


Figure 3: **Observation-selector and policy diagnostics.** Under the same backbone and evaluation budget, paired value supervision achieves the highest accuracy with the fewest acquired visual tokens. Post-scan decisions use all three actions rather than collapsing to a fixed fidelity. Appendix A gives counting rules and error rates.

9.0/7.0% to 6.0/4.0%. It matches or exceeds Always HIGH with 76.2/71.5% fewer tokens, demonstrating adaptive stopping and visual fidelity.

4.4 Case Study

MindSight follows neither a fixed resolution schedule nor a predetermined number of turns. In the first case, a person-specific headwear ambiguity prompts HIGH on [10–20s]; the focused evidence binds the attribute and immediately triggers STOP. In the second, LOW on [10–20s] resolves coarse lane occupancy but leaves finer alternatives open, so the updated policy shifts to [21–31s], escalates to HIGH, and then stops. The contrasting HIGH→STOP and LOW→HIGH→STOP paths demonstrate adaptive location, fidelity, and horizon. Full traces appear in Appendix D.

4.5 Efficiency Analysis

We distinguish newly acquired visual tokens from tokens reprocessed across calls, and report wall-clock serving measurements separately.

Processor-aware observation cost. For Qwen2.5-VL's patch size 14, temporal patch size 2, and spatial merge size 2 (Bai et al., 2025a), a processor-aligned clip with F frames and resolution $H \times W$

Table 3: A100-80GB inference efficiency (TP=4, batch size=1). TTFT is reported in seconds, TPOT in ms/token, and peak memory in GB/GPU.

Configuration	TTFT	TPOT	Memory (GB/GPU)
MindSight (prefix cache)	**3.46**	**28.4**	32.1
Uniform Sparse	4.72	29.1	**28.5**
MindSight (full replay)	7.92	30.8	41.2
Uniform Dense	16.90	29.7	64.2

produces

$$N_{\text{vid}} = \lceil F/2 \rceil \, (H/28)(W/28). \tag{12}$$

Thus a 64-frame 224^2 scan costs 2,048 tokens, while a 24-frame LOW/HIGH focus costs 768/3,072. For scan cost N_s and K focus clips with costs N_j, the acquired visual cost is simply $N_{\text{acq}} = N_s + \sum_{j=1}^{K} N_j$. One, three, and five full HIGH focuses therefore cost 5,120, 11,264, and 17,408 tokens, respectively. Appendix C reports the full processor-level accounting.

Replay can erase the saving. Without prefix reuse, a scan call, K full-history focus calls, and a separate answer call execute

$$N_{\text{stateless}} = (K+2)N_s + \sum_{j=1}^{K}(K-j+2)N_j \tag{13}$$

visual prefill tokens. For a 200-s video, five HIGH focuses acquire 17,408 tokens instead of 51,200 under uniform 2-fps sampling, yet stateless replay executes 75,776. Prefix caching can avoid much of this repeated prefill, making cache reuse and early stopping essential to realized efficiency.

A100-80GB inference efficiency. Table 3 reports serving efficiency with tensor parallelism 4 and batch size 1. TTFT includes all scan, tool, and reasoning calls before the first final-answer token; TPOT measures subsequent decoding. Prefix-cache inference reuses KV states for the identical multimodal prefix across turns. Full replay preserves the same browsing history and observations but recomputes the complete prefix at every turn. With 2.14 focus rounds on average, MindSight with prefix caching reduces TTFT by 26.7% relative to Uniform Sparse, with similar TPOT (28.4 vs. 29.1 ms/token), at 3.6 GB higher peak memory. Relative to Uniform Dense, it is 4.88× faster in TTFT, has similar TPOT (28.4 vs. 29.7 ms/token), and uses 50.0% less memory. Full replay increases TTFT from 3.46 to 7.92 s, confirming that cross-turn prefix reuse is important for inference efficiency. Appendix C.1 specifies the timing endpoints and controlled protocol.

5 Conclusion

We presented MindSight, framing long-video reasoning as cost-aware active perception. Paired Monte Carlo targets teach an adaptive scan–focus–reflect policy, while TB-GMPO optimizes complete browsing trajectories. Across two backbones and four benchmarks, our results show that efficient long-video reasoning benefits from learning where, how closely, and when to look rather than simply processing more frames.

Limitations

MindSight may miss brief or dispersed evidence under a fixed observation budget. Paired targets estimated with a frozen continuation policy can become stale, and multi-turn latency depends on cache reuse and deployment conditions. The current benchmark results cover multiple-choice QA; open-ended generalization remains untested. Future work will study adaptive budgets, target refresh, calibrated stopping, and efficient state reuse.

AI USE STATEMENT

In this work, we used generative AI tools for writing and editing parts of the experimental code, for analyzing the experimental data and computing statistics, and for polishing the language of this manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Boyu Chen, Zikang Wang, Zhengrong Yue, Kainan Yan, Chenyun Yu, Yi Huang, Zijun Liu, Yafei Wen, Xiaoxin Chen, Yang Liu, Peng Li, and Yali Wang. VideoChat-M1: Collaborative policy planning for video understanding via multi-agent reinforcement learning. *arXiv preprint arXiv:2511.19524*, 2026a.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. GRPO-CARE: Consistency-aware reinforcement learning for multimodal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026b. URL https://arxiv.org/abs/2506.16141.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling RL to long videos. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2507.07966.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2408.10188.

Zhe Chen et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2025c.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10k frames via hierarchical differential distillation. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2504.02438.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *arXiv preprint arXiv:2505.24718*, 2025.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-learned temporal focusing for long video reasoning. *arXiv preprint arXiv:2512.22315*, 2025. URL https://arxiv.org/abs/2512.22315.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.21776.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2405.21075.

Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. FrameMind: Frame-interleaved video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*, 2025. URL https://arxiv.org/abs/2509.24008.

Dong Guo et al. Seed1.5-VL technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Noriyuki Kugo, Tatsuya Ishibashi, Kosuke Ono, and Yuji Sato. VDMA: Video question answering with dynamically generated multi-agents. *arXiv preprint arXiv:2407.03610*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. URL https://arxiv.org/abs/2408.03326.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025b. URL https://arxiv.org/abs/2506.01908.

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*, 2025c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2311.17043.

Susan Liang, Chao Huang, Filippos Bellos, Jing Bi, Jason J. Corso, and Chenliang Xu. AdaTurn: Budget-aware test-time scaling for active visual perception agents. *arXiv preprint arXiv:2607.14547*, 2026. URL https://arxiv.org/abs/2607.14547.

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*, 2024a. URL https://arxiv.org/abs/2311.10122.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

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

Zhijian Liu et al. NVILA: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.04468.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2306.05424.

Chang Nie, Jiaju Wei, Junlan Feng, Chaoyou Fu, and Caifeng Shan. Light-Omni: Reflex over reasoning in agentic video understanding with long-term memory. *arXiv preprint arXiv:2607.05511*, 2026.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. LongVideo-R1: Smart navigation for low-cost long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Machel Reid et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.02051.

Michael S. Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Jongwoo Park, Kanchana Ranasinghe, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (BLIP-3-Video): You only need 32 tokens to represent a video even in vlms. *arXiv preprint arXiv:2410.16267*, 2025. URL https://arxiv.org/abs/2410.16267.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2410.17434.

Yan Shu, Liu Zheng, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2409.14485.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2307.16449.

Peng Wang, Shuai Bai, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Shihao Wang, Guo Chen, De-An Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jose M. Alvarez, Lei Zhang, and Zhiding Yu. VideoITG: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2026a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a. URL https://arxiv.org/abs/2406.08035.

Xiaodong Wang and Peixi Peng. Open-R1-Video. GitHub repository, 2025. URL https://github.com/Wang-Xiaodong1899/Open-R1-Video.

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*, 2024b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025b. URL https://arxiv.org/abs/2503.13377.

Yi Wang et al. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025c.

Ying Wang, Yanlai Yang, and Mengye Ren. LifelongMemory: Leveraging LLMs for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2024c.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis–verification multi-agent framework for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://arxiv.org/abs/2603.04977.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025d. URL https://arxiv.org/abs/2405.19209.

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. *arXiv preprint arXiv:2512.05774*, 2026c. URL https://arxiv.org/abs/2512.05774.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, 2024.

Caijun Yan, Yang Zhou, Meixing Shi, Haoran Sun, Yichen Li, Yuxiang Cai, and Yankai Jiang. REVEAL: A rubric-guided agent for explicit evidence sufficiency verification in long-video question answering. *arXiv preprint arXiv:2608.08612*, 2026a. URL https://arxiv.org/abs/2608.08612.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. InternVideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026b.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. SVAgent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. VideoARM: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. VideoExplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a.

Peiyuan Zhang et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. URL https://arxiv.org/abs/2410.02713.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.20673.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2406.04264.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A OBSERVATION-VALUE POLICY AND INFERENCE

The state contains the question, scan timeline, provisional answer, source-linked evidence, unresolved claims, and remaining budgets. A single policy chooses the interval and STOP/LOW/HIGH; the runtime only validates and executes. Evidence records distinguish support, contradiction, and insufficient visibility or coverage. Missing evidence does not automatically refute an option. Only actual tool observations create new visual evidence.

Algorithm 1 MindSight with observation-value decisions

Require: Video V, question Q, policy π_θ, maximum attempts $K = 5$
Require: Scan cap 64, focus cap 24, resolutions $224/448$, 1,024-token step limit
1: Scan at low resolution; initialize provisional answer $\tilde{y}$, evidence E, unresolved claims U, and $k = 0$
2: **loop**
3: Policy assesses $(Q, \tilde{y}, E, U)$ and selects (s, a), $a \in \{\text{STOP}, \text{LOW}, \text{HIGH}\}$
4: **if** $a = \text{STOP}$, $k = K$, or the generation limit is reached **then**
5: **break**
6: **end if**
7: $k \leftarrow k + 1$; validate interval, mode, and padded frame count
8: Execute a valid request; otherwise return a structured error
9: Policy updates $\tilde{y}$, source-linked evidence E, and unresolved claims U
10: **end loop**
11: Return $\tilde{y}$ and the stopping reason; failed examples remain in evaluation

Each attempted request consumes budget, including failures; a failed request that returns no frames adds zero acquired visual tokens. With $K = 0$ the agent answers after scanning. Requests cannot silently change semantic targets during validation. Coverage gaps can motivate a different interval, and unclear details can motivate HIGH. Redundant evidence may preserve a decision. A controlled test must fix all intervals and resolutions immediately after the scan, then compare with observation-dependent re-planning under matched budgets. Fixed-K HIGH alone is insufficient if later intervals still depend on observations.

Statistical boundaries. Action proportions divide STOP/LOW/HIGH counts by N_{valid}; LVBench has $N_{\text{valid}} = 3{,}492$ (1,549/1,135/808), and VideoMME-Long has $N_{\text{valid}} = 2{,}470$ (900/950/620). Unparsable actions are excluded; invalid-action rates are 1.99%/1.98%. A parseable action with an invalid interval remains in its action category, consumes an attempt, and contributes to the 1.12%/0.81% invalid-request rate. Mean decision rounds include the terminal STOP. Although the runtime enforces $K = 5$, no evaluated episode reached this forced cap. To compute false STOP, we run matched LOW/HIGH counterfactual branches offline on evaluable STOP states and test whether either improves value beyond τ_V; these diagnostic branches do not affect the reported inference trajectory or answer.

Open-ended questions. For multiple-choice QA, unresolved claims can correspond to candidate answers. For open-ended QA, the same state stores model-generated hypotheses and evidence gaps instead of option scores; STOP, LOW, and HIGH remain unchanged. This formulation separates the browsing policy from option elimination, but open-ended generalization remains untested by the multiple-choice results.

Training contract. Action-value branches are generated only on training videos under Equation 3. The current RL optimizer is TB-GMPO. XML validity is evaluated separately for each assistant turn and averaged to form the trajectory-level format reward. The loss includes assistant tokens, including action delimiters, and masks user/tool/visual/padding tokens. Empty padded turns have zero weight; old-policy probabilities, advantages, and action boundaries are detached. A constant-reward group has zero policy contribution but can retain a KL gradient. Reference-policy KL and global loss reduction are held fixed in optimizer comparisons. Tool outcomes, timestamps, processor grids, token counts, and stopping reasons are logged for reproducibility.

Table 4: Additional hyperparameters for SFT, observation-value supervision, and TB-GMPO.

Hyperparameter	Value
SFT implementation	
Distributed strategy	ZeRO-2
Per-device batch size	1
Gradient accumulation	2
Paired observation-value supervision	
Continuations per action M	32
Cost coefficient λ_C	0.10
Value margin τ_V	0.02
Action-loss weight λ_{act}	0.30
Bootstrap resamples	1,000
Simultaneous confidence	95% (Bonferroni)
Per-action two-sided interval	98.33%
TB-GMPO implementation	
Distributed strategy	ZeRO-3
Number of GPUs	4× A100-80GB
Per-device prompt batch	1
Gradient accumulation	1
Global prompt batch	4
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.95
Weight decay	0
Warmup	3% of total steps
LR schedule	Cosine decay
RL rollout and error handling	
Maximum tokens per step	1,024
Invalid-action reward	0
Timeout/tool-error reward	0

B TB-GMPO: Gradient Analysis and Numerical Checks

Assumptions. The sampled histories, old-policy log probabilities, advantages A_i, action boundaries, and weights $w_{ijt} = 1/(M_i L_{ij})$ are detached. All counted turns are nonempty; user/tool/visual/padding tokens are masked out. Let $J_i = A_i g_i$ be the policy term, with $|A_i| \le A_{\max}$, $c > 0$, and $0 < \epsilon < 1$. The reference KL term is excluded from the following bound.

Proposition. Define $C_i = 0$ for $A_i = 0$, $C_i = |A_i| \min(e^c, 1+\epsilon)$ for $A_i > 0$, and $C_i = |A_i| e^c$ otherwise. Then

$$\sum_{j,t} \left| \frac{\partial J_i}{\partial d_{ijt}} \right| \le C_i, \qquad \sum_{t} \left| \frac{\partial J_i}{\partial d_{ijt}} \right| \le \frac{C_i}{M_i}. \tag{14}$$

Proof. Since $u = c\tanh(d/c)$ lies in $(-c, c)$, the positive-advantage branch has $g_i \le \min(e^c, 1+\epsilon)$ and the negative branch has $g_i \le e^c$. Differentiation gives

$$\frac{\partial J_i}{\partial d_{ijt}} = A_i g_i w_{ijt} m_{ijt} \operatorname{sech}^2(d_{ijt}/c), \tag{15}$$

with clipping derivative $m_{ijt} \in [0, 1]$. Summing and using $\sum_{j,t} w_{ijt} = 1$ and $\sum_t w_{ijt} = 1/M_i$ proves both statements. If the additional assumption $\|\nabla_\theta \log \pi_\theta(x_{ijt} \mid h_{ijt})\| \le B$ holds, the chain rule yields $\|\nabla_\theta J_i\| \le C_i B$. The objective does not guarantee this Jacobian bound, lower variance, or convergence. □

Length balance and bias. With two turns of 8 and 128 tokens, uniform-token aggregation assigns 5.88% and 94.12% of its weight to them, whereas TB-GMPO assigns 50% to each. Actual gradients also depend on token Jacobians and clipping. Short noisy turns can be amplified. Because an action and its preceding reflection share one turn, equal turn weights do not prevent within-turn action-token dilution and do not provide step-specific credit assignment. The smooth guard approaches identity near zero and has vanishing derivatives for extreme ratios, which can suppress useful correction as well as instability.

Table 5: Controlled numerical analysis of surrogate-gradient coefficients. All cases use one assistant turn, $A = -1.5$, $\epsilon = 0.2$; TB-GMPO uses $c = 1.5$, $A_{\max} = 2$. The reported quantity is $\sum_t |\partial J/\partial d_t|$. One outlier sets one $d_t = 12$ and the rest to zero.

Log-ratio setting	Length	Sequence baseline	TB-GMPO
All zero	64	1.5	1.5
One outlier	8	6.7225	1.5832
One outlier	64	1.8093	1.5116
All three	64	30.128	0.45
All twelve	64	2.4413×10^{5}	3.0261×10^{-6}

The negative-advantage cases expose adverse upward ratio excursions, for which the sequence baseline's lower clipping branch remains inactive. The guard reduces coefficient magnitudes, while near-zero coefficients in the extreme case also illustrate saturation bias. The implementation computes each row and verifies the derivatives by finite differences (maximum absolute error below 6.8×10^{-10}); a broader numerical check covers 2,596 coordinates in 200 random cases, both advantage signs, empty-turn rejection, and per-turn bounds.

Relation to prior objectives. As $c \to \infty$, with uniform token weights, identical advantages, and matched token clipping thresholds, the policy term reduces to GMPO. With inactive clipping it also recovers the standard geometric sequence ratio. Geometric averaging alone is shared prior structure; TB-GMPO adds protocol-turn weights and the smooth guard while retaining GMPO-style token-level clipping. Paired observation utilities supervise actions separately and do not alter the shared-advantage derivation.

Observed training dynamics. Figure 4 complements the endpoint ablation with optimization-process evidence. TB-GMPO reaches 72.5% validation accuracy at 10k steps, exceeding GMPO and GRPO by 2.7 and 3.0 pp, respectively. Its absolute log-ratio p99 remains within 0.28–0.44 after initialization, whereas GMPO and GRPO exhibit repeated spikes up to 1.05 and 1.38. All methods show decreasing pre-clipping gradient norms as optimization progresses, but TB-GMPO remains lowest throughout and decreases more smoothly, ending at 3.3 versus 4.2/4.3 for GMPO/GRPO. Conditional KL rises early, peaks around 4k steps, and then declines toward a stable regime; TB-GMPO remains lowest throughout and ends at 0.019 nats/token, below GMPO (0.026) and GRPO (0.028). Together, these trajectories show that TB-GMPO attains higher validation accuracy with more controlled policy updates and less policy drift under the matched training setup.

C Additional Efficiency Accounting and Protocol

Count at the processor output. At patch size 14, temporal patch size 2, and spatial merge 2, a video clip costs $\lceil F/2 \rceil (H/28)(W/28)$ LLM visual tokens. The 64-frame 224^2 scan therefore costs 2,048; 24-frame LOW/HIGH clips cost 768/3,072, so choosing HIGH instead of LOW differs by 2,304 tokens. If LOW has already been processed, a subsequent HIGH request adds the full 3,072 tokens without feature reuse. Image-per-frame encoding, odd frame counts, resizing, and special tokens require their own accounting. Padding is applied within each clip, not once after concatenating all frames.

The last column assumes a scan call, K full-history focus calls, and a separate full-history answer call without prefix reuse. Answering in the existing call removes the last replay; retaining only text

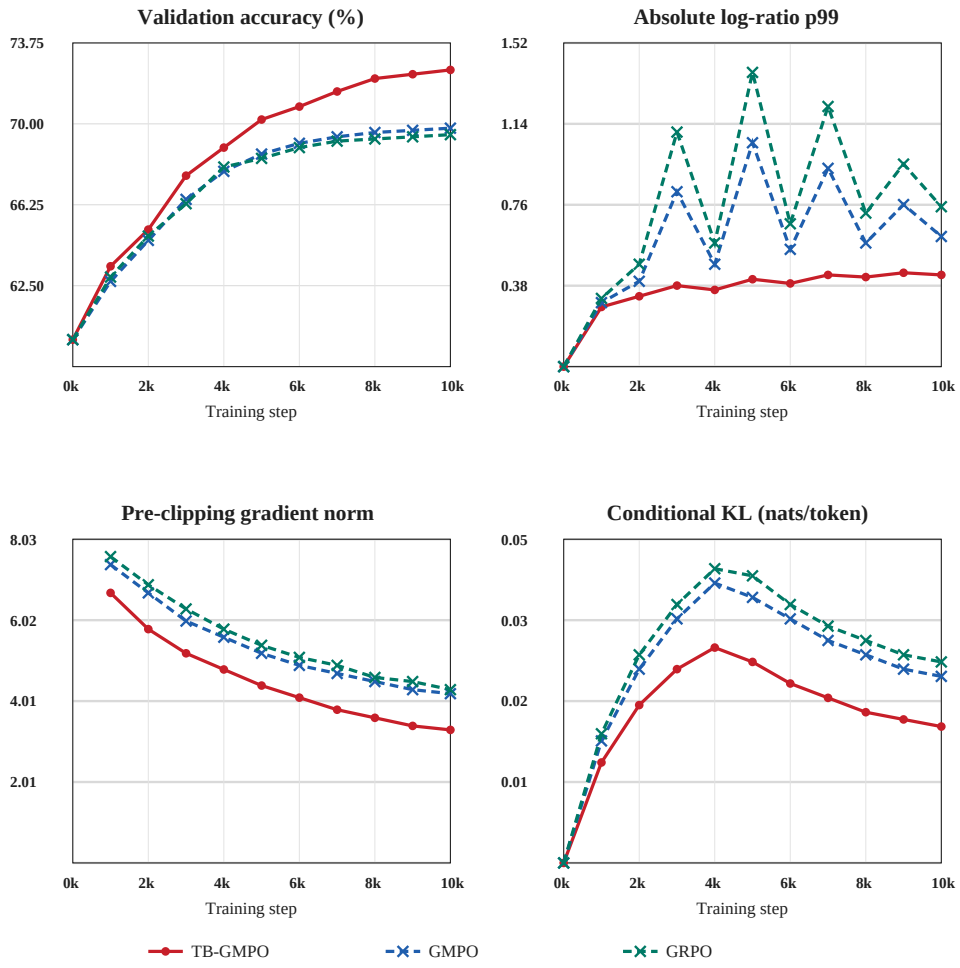


Figure 4: **Training dynamics under matched optimization settings.** Validation accuracy and conditional KL are means across runs at each displayed checkpoint. Absolute log-ratio p99 first computes the token-level p99 within each run and then averages across runs. Pre-clipping gradient norm reports the maximum across runs within each preceding 1,000-step window; it is undefined at step 0.

Table 6: Analytical acquired visual tokens after one scan and K full focus calls. LOW/HIGH share timestamps and differ in resolution. These counts do not measure answer quality or latency.

K	All LOW	All HIGH	HIGH replay + final
0	2,048	2,048	4,096
1	2,816	5,120	12,288
3	4,352	11,264	37,888
5	5,888	17,408	75,776

changes the cost and information available. With per-clip cost N_j and scan cost N_s,

$$N_{\text{acq}} = N_s + \sum_{j=1}^{K} N_j, \tag{16}$$

$$N_{\text{replay+final}} = (K+2)N_s + \sum_{j=1}^{K}(K-j+2)N_j. \tag{17}$$

Thus uniform 200 s video input (51,200 tokens) can cost less than five HIGH focus calls with full replay (75,776), despite fewer acquired tokens. A mixed-resolution policy must be evaluated from its actual request sequence; LOW is cheaper only if it preserves information needed for the answer.

Adaptive stopping affects the cost distribution. For fixed successful focus cost N_f and random count K, before a separate answer call,

$$\mathbb{E}[N_{\text{acq}}] = N_s + N_f\mathbb{E}[K], \tag{18}$$

$$\mathbb{E}[N_{\text{replay}}] = N_s(1+\mathbb{E}[K]) + \tfrac{1}{2}N_f(\mathbb{E}[K^2]+\mathbb{E}[K]). \tag{19}$$

Always taking three HIGH calls and taking one/five calls with equal probability both give $\mathbb{E}[K]=3$ and 11,264 acquired tokens, but expected replay is 26,624 versus 32,768. This calculated example shows why mean frame or call count is insufficient. Variable clip sizes, failures, and overlap need the general per-call ledger.

Cost and accuracy criteria. For every matched example, log acquired visual tokens, executed prefill tokens, generated tokens, and request-to-answer-completion latency. Executed prefill must be distinguished from cached logical prompt length. Report processor/model revision, hardware, load, cache state, timeouts, and peak memory. Token counts are not a linear latency model. An accuracy-preserving gain requires a predeclared non-inferiority tolerance, a supported lower bound on adaptive-minus-baseline accuracy, and lower measured cost. Bootstrap source videos with all their questions, pair methods on identical clusters, and report seed variation separately.

C.1 A100 Inference Protocol and Compute Accounting

TTFT denotes the time from receipt of the video–question request to the first final-answer token, including preceding scan, tool, and reasoning calls. TPOT is the mean interval between subsequent output tokens. Peak memory denotes the maximum live tensor allocation per GPU, distinct from reserved memory and preallocated KV-cache capacity. Total training compute is 362.4 A100 GPU-hours, computed as $\sum_r n_{\text{GPU},r} t_r$ over trajectory generation, utility branching, SFT, and TB-GMPO. Focus counts include failed attempts.

Controlled evaluation scope. We compare observation-dependent re-planning with fixed-plan inference under matched budgets, and use fixed LOW/HIGH and confidence stopping to isolate resolution and termination decisions. All agent comparisons use the same backbone, feasible context, modality access, and budget. Optimizer comparisons hold the reward, reference policy, KL estimator, and global loss reduction fixed.

D Case Studies: From Evidence Gaps to Actions

The following cases illustrate how an evidence gap leads to candidate intervals, a value-guided request, an observation update, and a final answer. H1/H2 and R1/R2 identify the evidence used at each step. The resulting HIGH→STOP and LOW→HIGH→STOP trajectories show that the policy adapts its location, fidelity, and stopping time to the evidence acquired so far rather than following a fixed action schedule. The displayed policy decisions are single-action outputs; option-support values are evidence-assessment scores rather than calibrated probabilities.

Trace details. The St. Patrick's Day illustration in Figure 1 displays video timestamps, while the other cases distinguish requested and decoded intervals. The runtime enforces the 24-frame sampling cap within each decoded temporal window. Each displayed decision is produced by one policy

pass; inference does not run counterfactual Monte Carlo branches or a separate value network. Support scores are evidence-assessment signals rather than calibrated probabilities, and unseen national identifiers do not prove absence.

D.1 Local ambiguity: identifying the queried person's headwear

The question refers to one man in a scene containing another person with a red bandana. A global scene description is therefore insufficient: the agent must bind the headwear observation to the correct person. The scan identifies a dining interaction around 10–20 s, then requests a closer look. Figure 5 shows how the focused observation supplies the missing attribute and distinguishes the two people.

Adaptive behavior. The unresolved person-specific attribute requires fine visual detail, so the policy selects HIGH directly rather than first spending a turn on LOW. Once the focused observation binds the headwear to the correct person, the updated state supports STOP. The requested and decoded intervals are recorded separately to preserve the exact tool interaction.

D.2 Comparative uncertainty: checking lane occupancy and alternatives

The second question requires comparing four claims about a race. The initial scan identifies the event but cannot reliably resolve lane numbers, shoes, or nationality indicators. A LOW observation first resolves the coarse lane-occupancy cue, after which a HIGH observation inspects finer shoe and nationality cues. The distinction between direct support and incomplete refutation is central to the case.

Adaptive behavior. The action sequence is observation-dependent rather than fixed. The scan first triggers LOW on [10–20s] because lane occupancy is a coarse cue. Conditioned on R1, the policy shifts to [21–31s] and escalates to HIGH for finer shoe and nationality cues; conditioned on R2, it then selects STOP. The policy therefore adapts where to look, at what fidelity, and for how many turns while preserving unresolved claims.

Case 1: Question and initial uncertainty

Question. What decoration is the man wearing on his head while sitting at the table and talking with the American man?
Options. A. A top hat; B. A headscarf; C. He isn't wearing anything; D. A duckbill cap.
Ground truth / prediction: D / D.

1. Global Scan (LOW, full video)

<action> SCAN </action>
<caption> Dining interaction followed by food preparation and a map of Pakistan. </caption>
<timeline> red-bandana man at table [0–10s]; two men conversing [10–20s]; food preparation and map [20–40s]. </timeline>
<candidates> [0–10s]: nearby bandana; [10–20s]: target conversation. </candidates>
<unresolved> Which headwear belongs to the queried man? </unresolved>
<think> The scan can confuse the target with the nearby bandana wearer; [10–20s] is the candidate that binds identity to headwear. </think>
<reason> Inspect the interval containing the target conversation rather than the distractor seen earlier. </reason>

2. Value-Guided Focus

<candidate> requested [10–20s]; decoded [9–21s] </candidate>
<action> HIGH on [10–20s] </action>
<reason> Resolving person identity and cap shape requires high spatial fidelity; the expected gain justifies HIGH's cost. </reason>
<observation> H1: the queried man wears a blue cap; H2: the other man wears a red bandana. </observation>

3. Reflect: Update Answer Support

<evidence> H1 binds the blue cap to the target; H2 separates the distractor. The cap shape matches a duckbill cap. </evidence>
<option-support> A:0.0, B:0.0, C:0.0, D:0.9 </option-support>
<action> STOP </action>
<reason> The target-specific observation resolves the question; further observation is not worth its cost. </reason>

4. Stop and Answer

<answer> D </answer>
<reason> The high-resolution observation supports D through target-person identification and headwear classification, rather than the generic presence of a cap elsewhere in the scene. </reason>

Figure 5: **Targeted evidence acquisition.** Faced with fine-grained identity uncertainty, the policy selects HIGH directly and stops once the focused evidence resolves the question. H1/H2 identify the acquired evidence.

Case 2: Question and comparison targets

Question. Which statement is true for Men's 100M Round 1 Heat 1?
Options. A. No athlete is on Track 1; B. The athlete with green running shoes is on Track 7; C. Two athletes have the same nationality; D. The athlete from USA is on Track 9.
Ground truth / prediction: A / A.

1. Global Scan (LOW, full video)

<action> SCAN </action>
<caption> Stadium introduction, starting blocks, and the 100M race. </caption>
<timeline> crowd [0–10s]; runners in starting blocks [10–20s]; race and lane views [20–40s]. </timeline>
<candidates> [10–20s]: lane occupancy; [21–31s]: shoes, uniforms, and nationality cues. </candidates>
<unresolved> Lane 1 occupancy; Track 7 shoes; shared nationality; USA athlete on Track 9. </unresolved>
<reason> Inspect the starting-block interval first because it can directly resolve lane occupancy. </reason>

2. First Value-Guided Focus

<candidate> requested [10–20s]; decoded [9–21s] </candidate>
<action> LOW on [10–20s] </action>
<reason> The coarse lane-occupancy cue is visible at LOW fidelity; fine-grained alternatives remain unresolved. </reason>
<observation> R1: Track 1 appears empty, providing direct support for option A. </observation>
<option-support> A:0.5, B:-0.3, C:0.0, D:0.0 </option-support>

3. Second Value-Guided Focus

<candidates> [10–20s]: revisit lane occupancy; [21–31s]: inspect remaining alternatives. </candidates>
<action> HIGH on [21–31s] </action>
<reason> R1 supports A, but B–D still depend on fine shoe and nationality cues in [21–31s]. </reason>
<observation> R2: dark uniform on Track 7; green uniform on Track 6; visible flags; no visible USA identifier on Track 9. </observation>
<option-support> A:0.8, B:-0.5, C:-0.5, D:-0.3 </option-support>

4. Reflect: Preserve Evidence Limits

<evidence> R1 directly supports A through empty-lane evidence. </evidence>
<unresolved> Uniform color does not prove shoe color; partial flags do not settle all nationalities; a missing USA identifier is not proof of absence. </unresolved>
<action> STOP </action>
<reason> A has direct support, whereas another costly observation is unlikely to overturn the decision. </reason>

5. Stop and Answer

<answer> A </answer>
<reason> The agent selects A from the observed lane occupancy and does not overstate the incompletely checked alternatives. </reason>

Figure 6: **Adaptive evidence acquisition.** Conditioned on updated evidence, the policy uses LOW for coarse lane occupancy, shifts interval and escalates to HIGH for finer alternatives, and then selects STOP. R1/R2 make the evidence dependencies explicit; option-support values remain heuristic scores.