CALLING IS NOT LOOKING: ENTROPY-RESOLVED EVIDENCE ACQUISITION FOR VIDEO AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

MLLM-based video agents call tools to retrieve missing visual evidence, but a call does not guarantee the evidence is actually or properly used. To address this, we propose **E**ntropy-**R**esolved Evidence **A**cquisition (ERA), an entropy-guided learning framework. ERA tracks the agent's top high-entropy tokens, which mark where the agent is most uncertain how to continue, and checks whether this uncertainty resolves only after evidence arrives. It uses this signal for both failure diagnosis and reinforcement learning. For diagnosis, ERA reveals two distinct failure modes: agents trained by RL with single or combined tool-calling rewards tend to commit to an answer before the tool call returns, which means they **overlook** the tool returns, while those cold-started under LLM-as-a-Judge supervision often answer after a single tool return using templated reasoning even with uninformative evidence, which means they **overrely** on the tool returns. We then translate this diagnosis into an **entropy-change** RL reward that scores where the agent commits and contrasts its reasoning under uninformative and decisive returns, encouraging the agent to autonomously re-seek while the high-entropy token rate remains elevated after an uninformative return and to commit as this rate drops after decisive evidence arrives. As a result, ERA achieves significant performance gains on four video reasoning benchmarks, with notable improvements in evidence sensitivity and active tool calling. Taken together, our results show that an agent's intrinsic high-entropy token distribution carries rich and important information, opening up a promising new direction for evidence-aware tool use in video agents.

1 INTRODUCTION

When answering questions about videos, humans do not attend to every moment equally; instead, they skim the content and selectively revisit decisive events. Current multimodal large language models (MLLMs), however, still struggle to revisit a video in this way. Limited by their context length, they observe a video through a fixed budget of uniformly sampled frames, which routinely miss short, decisive events. Question-conditioned frame selectors identify more relevant frames (Yu et al., 2024; Tang et al., 2025), but they typically select frames in a single pass before reasoning begins and rarely revisit the selection when it misses critical evidence. To address this limitation, MLLM-based video agents interleave reasoning over a coarse view with tool calls that retrieve the moments they need. Prior work trains these agents with reinforcement learning (RL) on outcome rewards and reports more frequent tool calls and higher accuracy (Ding et al., 2025; Yang et al., 2025; Gao et al., 2026).

However, calling a tool does not guarantee that the returned evidence is actually or properly used. For image agents, swapping the returned crops for random or blank images changes far fewer answers than the agents' accuracy gains would suggest (Wang et al., 2026b), and most tool-solved problems can also be solved without the tool (Guo et al., 2026). For video agents, the first call may itself miss the decisive moment and return an uninformative segment. The agent must then notice from the returned content that it has not yet seen what it needs and re-seek, which is what video agents were introduced to do. Current training signals do not tie such re-seeking to what the returned frames show: an outcome reward evaluates only the final answer, tool-calling rewards credit the act of calling, an LLM-as-a-Judge scores the written reasoning, and evidence-grounded

Code and the trained model are released anonymously on Hugging Face under the name *era-zoomer*.

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.

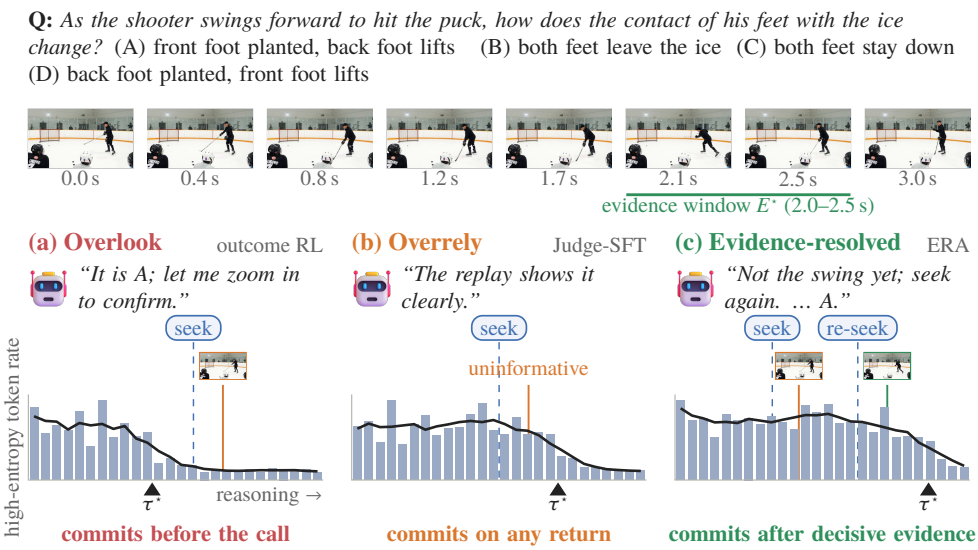

Figure 1: **Where an agent still reasons reveals whether it looked.** Top: a tool-required question with its eight coarse frames, which miss the evidence window. Bottom: schematic traces of the three behaviors; bars show the high-entropy token rate (top 20% of the trajectory) along the reasoning and the black line its trend, blue dashed lines mark tool calls, orange and green marks the uninformative and decisive returns, and ▲ the commitment point $\tau^\star$. (a) *Overlook*: the rate drops before the call, so the return has little influence on the answer. (b) *Overrely* (Judge-SFT on another question): the rate stays high until the return and drops right after it although the return is uninformative. (c) *Evidence-resolved* (ERA): the rate stays high after an uninformative return, the agent re-seeks, and it drops only after decisive evidence. Population statistics are in Figure 4.

rewards (Huang et al., 2026; Xu et al., 2025) measure sensitivity to regions already present in the input. None of these signals measures whether the returned evidence informs the agent's reasoning.

In this paper, we propose *Entropy-Resolved Evidence Acquisition* (ERA), a framework that trains a video agent to commit to an answer only after decisive evidence has arrived. ERA measures evidence use through the agent's top high-entropy tokens: the minority of tokens where the agent is most uncertain how to continue, which mark the decision forks of a reasoning chain (Wang et al., 2025b; Pan et al., 2026). We locate the *commitment point*, the step at which the high-entropy token rate drops (Figure 1), and check whether this uncertainty resolves only after decisive evidence arrives. The position of the commitment point relative to the tool calls distinguishes three behaviors that accuracy and call counts miss: the agent may commit before the call (*overlook*), on any return (*overrely*), or only after decisive evidence (*evidence-resolved*). The same signal serves both as a diagnostic and as an RL reward.

As a diagnostic, the commitment point of ERA reveals two opposite failure modes in existing training recipes. Agents trained by RL with tool-calling rewards tend to **overlook** the tool returns: DeepEyes-7B, for example, commits to an answer before the call returns in 63% of trajectories and gives the same answer for 97% of questions, whether the return is real, random, gray, or the evidence window. Agents cold-started under LLM-as-a-Judge supervision instead tend to **overrely** on the returns: our Judge-SFT baseline answers after a single return with templated reasoning, even when that return is uninformative, and rarely re-seeks. Conventional signals do not correct either failure: an outcome reward with a call cost pushes the agent toward not calling tools, and an LLM-as-a-Judge saturates once the agent learns templated explanations such as "the returned frames show … therefore". To address both failures, ERA turns the commitment point into a reward with two terms. The *location term* scores the position of the commitment point and credits the committing turn, penalizing commitment before the call. The *contrastive term* compares continuations of the same prefix under an uninformative and a decisive return, encouraging the agent to re-seek while its high-entropy token rate remains elevated after an uninformative return and to commit once the rate drops after decisive evidence arrives.

We evaluate ERA on tool-required questions from VideoNet-Gen, CG-Bench, NExT-GQA, and Video-MME, and separate the roles of supervised fine-tuning (SFT) and RL. We construct two SFT sets: single-turn trajectories that answer correctly, and counterfactual multi-turn trajectories that first receive an uninformative return and then re-seek the evidence window. Both SFT variants reach similar accuracy (0.47 vs. 0.48), and the counterfactual variant raises re-seek to 0.39. This re-seek, however, collapses to 0.04 under a pressure prompt asserting that the return is exactly what the agent requested, and the agent reverts to overrely. RL with the ERA reward is the stage that sustains re-seek under this pressure: RL alone reaches 0.33, and combining both SFT sets with RL raises it to 0.40 at comparable accuracy, whereas outcome and judge rewards from the same initialization reach only 0.12–0.21. These results indicate that SFT teaches the form of re-seeking, whereas the entropy-change reward ties it to what the returned frames show. The main contributions are summarized as follows:

- We locate the commitment point in the agent's own high-entropy tokens, which measures evidence use without a judge or a second model, and use it to expose two opposite failures of current training recipes: overlook and overrely.
- We propose ERA, which turns the commitment point into a turn-level entropy-change reward that combines a graded location term with a contrastive term, rewarding the agent for committing only after decisive evidence arrives.
- We separate the roles of SFT and RL, showing that SFT on counterfactual re-seek trajectories teaches the form of re-seeking but collapses under a pressure prompt, whereas RL with the ERA reward sustains re-seek under the same pressure at comparable accuracy.

2 Related Work

Multimodal reasoning with RL and tools. Reinforcement learning (RL) with verifiable rewards elicits long reasoning in both language (Guo et al., 2025; Shao et al., 2024; Yu et al., 2025) and video domains (Feng et al., 2025). A parallel line puts tools inside the reasoning chain: cropping and zooming in DeepEyes and DeepEyesV2 (Zheng et al., 2025; Hong et al., 2025), pixel-space operations in Pixel Reasoner (Wang et al., 2025a), and search in MMSearch-R1 (Wu et al., 2025). Audits question what these tools contribute: swapping returned crops for random or blank images changes few answers (Wang et al., 2026b), while most tool-solved problems can be resolved without tool calls (Guo et al., 2026). FaithEyes (Wang et al., 2026a) scores tool use with a judge, Faithful-MR1 (Tian et al., 2026) anchors attention to evidence in the image, and faithfulness tests perturb the chain of thought (Turpin et al., 2023; Lanham et al., 2023). We share the counterfactual intervention of these audits but apply it to evidence the agent must *acquire*, measure its effect in the agent's own tokens, and train on it.

Video understanding and video agents. Beyond frame selection (Yu et al., 2024; Tang et al., 2025) and longer contexts trained by RL (Chen et al., 2025), video agents retrieve evidence while reasoning. Some prompt proprietary language models (Zhang et al., 2025; Yuan et al., 2025; Wang et al., 2025d). Others train open base models to perform temporal cropping: VideoZoomer (Ding et al., 2025), LongVT (Yang et al., 2025), Video-Thinker (Wang et al., 2025c), VideoTemp-o3 (Liu et al., 2026b), FrameThinker (He et al., 2025), VideoSearcher (Gao et al., 2026), and open-world video agents (Liu et al., 2026a). They are trained with outcome and tool-calling rewards after an SFT cold start and evaluated by accuracy and call statistics. We test whether such agents re-seek when the returned content calls for it, and separate what SFT and RL each contribute to this behavior. Evidence-grounded training (Huang et al., 2026; Xu et al., 2025) rewards sensitivity to regions already in the input, and step-level rewards from verifiers or judges (Lightman et al., 2023; Zheng et al., 2023) score text; neither sees whether an acquired return changed the reasoning.

Token entropy in RL for reasoning. Token entropy is central to RL for reasoning models (Taricco, 2026). Because policy entropy collapses early while performance saturates, prior work proposes clip-higher (Yu et al., 2025) and covariance-based clipping of the tokens that drive the collapse (Cui et al., 2025), and analyzes how clipping, off-policy updates and data diversity shape entropy (Jin et al., 2025). At the token level, a minority of high-entropy "forking" tokens carries most of the benefit of RL (Wang et al., 2025b). High-entropy regions mark pivotal steps and can

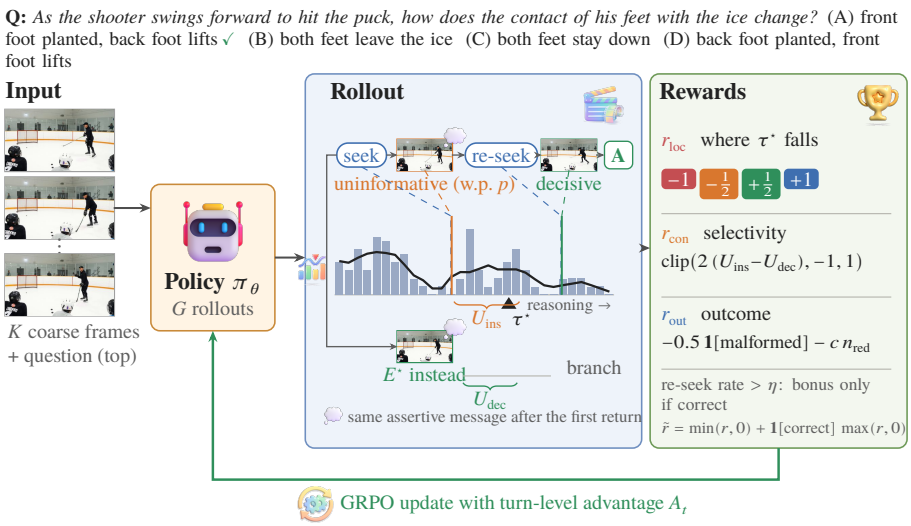

Figure 2: **ERA training loop** on the question of Figure 1. The agent calls a temporal tool in an environment where, with probability p, the first return is swapped for an uninformative one that keeps the requested time labels; decisive or not, the first return is followed by the same assertive message. The entropy trace (the real ERA trajectory of Figure 3; bars: high-entropy token rate per token bin; line: trend) gives the commitment point $\tau^\star$. The position of $\tau^\star$ relative to the calls and the first decisive return sets the location term, and comparing the continuation after the uninformative return with the same prefix continued under $E^\star$ sets the contrastive term; both feed the GRPO update. U_{ins} and U_{dec} are the high-entropy token rates after the uninformative return and after $E^\star$ in the branch.

shape the advantage (Cheng et al., 2025). PEA (Pan et al., 2026) aligns their positions with prototypes, and in visual reasoning entropy alone misses perception-critical tokens (Jin et al., 2026). These methods use entropy to decide *how strongly* to update each token. ERA instead locates the agent's commitment point from *where* its high-entropy tokens fall relative to a tool call and its return. It rewards that position rather than any entropy level, and it contrasts two returns to cancel uncertainty that does not depend on the returned content.

3 Entropy-Resolved Evidence Acquisition

We train a video agent to commit to an answer only after decisive evidence has arrived (Figure 2). ERA judges a trajectory not by what the agent writes about a tool return, but by whether the agent's decision was still open when the return arrived. It locates the commitment point in the agent's own high-entropy tokens and turns its position relative to the tool calls into a turn-level RL reward.

3.1 Problem Setup

The agent receives a question with its options and observes the video through K uniformly sampled low-resolution frames, which we call the *coarse frames*. To acquire missing evidence, it can call *video_seek*(s, e, n), which returns n frames sampled from the segment $[s, e]$, or *image_zoom*(i, box), which crops the region box from coarse frame i. A call issued in turn t yields a *return*, the frames shown at the start of turn $t{+}1$; a trajectory has at most $T_{\max}$ turns, each ending with a call or the answer. The *evidence window* $E^\star$ of a question is the short segment in which the decisive event occurs. A question is *tool-required* if it cannot be answered from the coarse frames but can be once frames from $E^\star$ are added; a *control* question can be answered from the coarse frames alone. A return is *decisive* if the segment it covers overlaps $E^\star$ and *uninformative* otherwise. Since the agent must locate $E^\star$ from the coarse frames on its own, a counterfactual return, which swaps the real frames for an uninformative segment, tests whether the agent's calls respond to what it is missing.

3.2 High-Entropy Tokens and Commitment

Entropy trace. The next-token entropies $h_1, \ldots, h_T$ of the reasoning tokens are concatenated over turns, excluding tokens inside tool-call and answer tags. Following PEA (Pan et al., 2026), we set $z_i = 1$ if h_i is among the highest ρ fraction of entropies in the trajectory and $z_i = 0$ otherwise; the binary sequence z is the *entropy trace*. The *high-entropy token rate* of a span is $U(a, b) = \frac{1}{b-a} \sum_{i=a}^{b-1} z_i$. For each turn, the rate is computed over its first L reasoning tokens only, so that turn length does not affect it.

Commitment. A high rate U indicates that the agent is still choosing between continuations; a sustained drop in U indicates that it has settled on one and is now elaborating on it (Figure 1). We therefore define commitment by this drop, using the entropy trace alone.

3.3 Detecting the Commitment Point

We locate the drop by splitting the entropy trace into two segments, each with its own high-entropy token rate, and choosing the split that best fits the trace. This is the maximum-likelihood estimate of a single change point in a binary sequence (Hinkley & Hinkley, 1970):

$$\tau^\star = \arg\max_c \ \Big[\ell\big(z_{1:c}; \hat{p}_1\big) + \ell\big(z_{c+1:T}; \hat{p}_2\big)\Big] \quad \text{s.t. } \hat{p}_2 \le \kappa\, \hat{p}_1, \tag{1}$$

where ℓ is the Bernoulli log-likelihood of a segment at its empirical rate $\hat{p}$, and the constraint admits only splits after which the rate falls to at most $\kappa < 1$ times its earlier value. The split $\tau^\star$ is the *commitment point*. We accept it only if the two-segment fit improves on a single segment by a log-likelihood ratio of at least λ and each segment contains at least m tokens; otherwise the trajectory has *no commitment*.

Three behaviors. The position of the commitment point relative to the calls and returns defines three behaviors. Let k be the turn issuing the first successful call, j the turn issuing the first call whose return is decisive, and $t(\tau^\star)$ the turn containing $\tau^\star$. A trajectory on a tool-required question is

$$\text{kind} = \begin{cases} \textit{overlook} & t(\tau^\star) \le k, \text{ or answered without any call,} \\ \textit{overrely} & t(\tau^\star) > k \text{ and no decisive return by turn } t(\tau^\star), \\ \textit{evidence-resolved} & t(\tau^\star) > j, \\ \textit{none} & \text{no commitment point.} \end{cases} \tag{2}$$

Evidence-resolved trajectories are those in which the agent uses the evidence it acquires: its decision is still open when a return arrives and is made only after a decisive return.

3.4 Entropy-Resolved RL

Training environment. In a fraction p of the rollouts on tool-required questions, the first return is replaced by a counterfactual return, a segment of the same width outside $E^\star$ labeled with the requested times. Every first return is followed by the same assertive message stating that the frames are exactly what the agent requested. A re-seek returns exactly the requested segment, and the prompt does not mention re-seeking. Each swapped rollout is also branched at the swapped return with frames from $E^\star$; the branch is used only to compute the contrastive term.

Reward. The reward consists of a location term, a contrastive term, and an outcome term (Table 1). The location term r_{loc} scores where the commitment point falls, from -1 for overlook to $+1$ for evidence-resolved. The contrastive term compares the two continuations of a swapped rollout,

$$r_{\text{con}} = \text{clip}\big(2\,(U_{\text{ins}} - U_{\text{dec}}), -1, 1\big), \tag{3}$$

where U_{ins} and U_{dec} are the high-entropy token rates of the turn after the uninformative return and of the corresponding turn in the branch. The two continuations share the prefix and the message and differ only in the returned frames; we call $U_{\text{ins}} - U_{\text{dec}}$ the *selectivity* of the rollout. Once the re-seek rate of the training rollouts exceeds η, the positive part of each process term is paid only on correct answers:

$$\tilde{r} = \min(r, 0) + \mathbf{1}[\text{correct}]\ \max(r, 0), \qquad r \in \{r_{\text{loc}}, r_{\text{con}}\}. \tag{4}$$

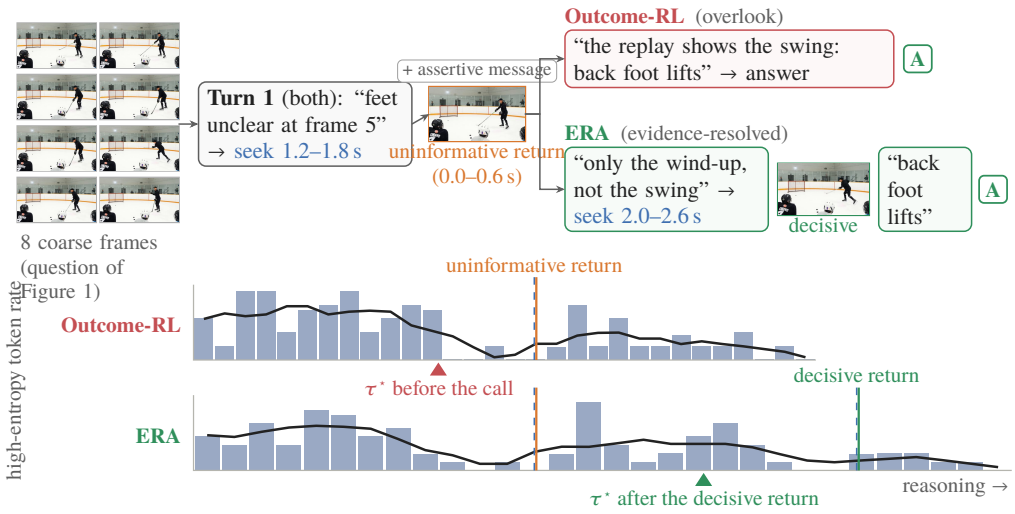

Figure 3: **Two agents, one question, one uninformative return** (the question of Figure 1). Both agents start from the same initialization and make the same first call, whose return is swapped for a segment outside the evidence window. The outcome-trained agent reads a swing into frames that contain none and answers; its commitment point lies before the call (overlook). The ERA agent notes that the swing is missing, re-seeks, and commits only after the decisive return (evidence-resolved). Both answers are correct, so accuracy does not separate them. Quoted reasoning is translated from the agent's Chinese output.

Table 1: **Reward terms.** The location term is credited to the turn containing the commitment point; the contrastive term compares the same prefix under an uninformative and a decisive return. After the re-seek rate exceeds η, the process terms pay their positive values only on correct answers.

Term	Trajectory	Value
r_{loc}	overlook, or answered without calling	-1
	overrely	$-\frac{1}{2}$
	re-seek after an uninformative return without committing	$+\frac{1}{2}$
	evidence-resolved	$+1$
r_{con}	swapped rollout and its $E^\star$ branch	$\text{clip}\big(2\,(U_{\text{ins}} - U_{\text{dec}}), -1, 1\big)$
r_{out}	every rollout	$-0.5\,\mathbf{1}[\text{malformed}] - c \cdot n_{\text{red}}$

The outcome term penalizes malformed outputs and charges a cost c for each redundant call, that is, each call after a decisive return and any call on a control question.

Optimization. Each term is standardized within the G rollouts of a question, as in GRPO (Shao et al., 2024), with A_{con} standardized among swapped rollouts only. Process credit goes only to the turn that earned it:

$$A_t = A_{\text{out}} + w\,\big(\mathbf{1}[t = t(\tau^\star)]\,A_{\text{loc}} + \mathbf{1}[t = k{+}1]\,A_{\text{con}}\big), \tag{5}$$

where w weights the process terms, $t(\tau^\star)$ is the committing turn, and $k{+}1$ is the turn that decides whether to re-seek. The policy is updated with a clipped policy-gradient objective (Schulman et al., 2017).

4 Experiments

4.1 Setup

Models and baselines. The policy is Qwen3.5-9B (Qwen Team, 2026). We compare with two external agents, DeepEyes-7B (Zheng et al., 2025) and VideoZoomer (Ding et al., 2025), each run with its own backbone and tool format, and with four recipes on our backbone: Judge-SFT, two Outcome rewards, and MM-Judge, the last three trained with RL from SFT-Mix.

Table 2: **Training hyperparameters.**

Hyperparameter	Value	Hyperparameter	Value
SFT epochs	3	RL learning rate	1e-6
SFT learning rate	1e-3	Clip range	0.2 / 0.28
Rollouts per question G	8	KL coefficient	0.01
Questions per round	240 + 120	Temperature	0.9

Table 3: **Main results.** Acc.: real returns and plain messages. Acc. P: first return swapped for a counterfactual one and assertive messages after up to three returns. Hal. P: share of swapped trajectories classified as overrely by Eq. 2 under the same protocol; models that commit before calling, such as DeepEyes-7B, overlook the return and thus have low Hal. Small numbers give the change from the row's start point (VideoZoomer, SFT-Mix, or Qwen3.5-9B). Bold: best among the trained models on our backbone. Video-MME has no evidence windows.

	VideoNet-Gen			CG-Bench			NExT-GQA		Video-MME
Method	Acc. ↑	Acc. P ↑	Hal. P ↓	Acc. ↑	Acc. P ↑	Hal. P ↓	Acc. ↑	Acc. P ↑	Acc. ↑
Qwen3.5-9B	0.49	0.33	0.02	0.22	0.18	0.03	0.62	0.55	0.25
DeepEyes-7B	0.35	0.36	0.08	0.23	0.23	0.05	0.65	0.61	0.33
VideoZoomer	0.36	0.22	0.55	0.22	0.14	0.61	0.57	0.41	0.28
VideoZoomer+ERA	0.37 ↑.01	0.37 ↑.15	0.14 ↓.41	0.23 ↑.01	0.22 ↑.08	0.19 ↓.42	0.58 ↑.01	0.56 ↑.15	0.29 ↑.01
Judge-SFT	0.49	0.30	0.47	0.25	0.17	0.52	0.61	0.50	0.20
SFT-Mix	0.49	0.31	0.41	0.26	0.19	0.44	0.58	0.47	0.27
+ Outcome (per-call cost)	0.52 ↑.03	0.34 ↑.03	0.38 ↓.03	0.26 =	0.20 ↑.01	0.40 ↓.04	0.61 ↑.03	0.49 ↑.02	0.22 ↓.05
+ Outcome (redundant-call cost)	0.52 ↑.03	0.36 ↑.05	0.35 ↓.06	0.25 ↓.01	0.21 ↑.02	0.37 ↓.07	0.56 ↓.02	0.48 ↑.01	0.25 ↓.02
+ MM-Judge	0.50 ↑.01	0.33 ↑.02	0.44 ↑.03	0.26 =	0.19 =	0.46 ↑.02	0.60 ↑.02	0.50 ↑.03	0.24 ↓.03
+ **ERA**	**0.53** ↑.04	**0.50** ↑.19	0.12 ↓.29	**0.27** ↑.01	**0.25** ↑.06	0.16 ↓.28	**0.62** ↑.04	**0.58** ↑.11	**0.28** ↑.01
ERA-RL	0.44 ↓.05	0.40 ↑.07	**0.05** ↑.03	0.21 ↓.01	0.19 ↑.01	**0.06** ↑.03	0.57 ↓.05	0.52 ↓.03	0.23 ↓.02

Training data. We generate multiple-choice questions with Gemini (Gemini Team et al., 2023) on 724 VideoNet (Yadav et al., 2026) videos, filter them manually (VideoNet-Gen), and keep 409 tool-required questions, which Qwen3.5-9B answers correctly in at most one of four samples from the coarse frames but in at least three of four with frames from $E^\star$, plus 409 control questions. For SFT, SFT-Single contains 1,109 trajectories that answer correctly with at most one call, SFT-CF contains 709 demonstrations in which a teacher re-seeks $E^\star$ after an uninformative first return, and SFT-Mix is their union.

Training settings. Table 2 lists the training hyperparameters. Training runs on 8 GPUs with 80 GB of memory.

Evaluation. Evaluation uses certified tool-required questions and greedy decoding: held-out VideoNet-Gen videos (100) in domain and, for zero-shot transfer, CG-Bench (400; Chen et al., 2024) and NExT-GQA (84; Xiao et al., 2024), whose annotated clue intervals and answer spans serve as $E^\star$, and Video-MME (169; Fu et al., 2024).

4.2 Main Results

Real returns. With real returns, the recipes on our backbone differ little: ERA is the most accurate on all four benchmarks, but by only 0.01 in each case (Table 3).

Under pressure. Swapping the first return and adding assertive messages lowers accuracy for every recipe, but much less under the ERA reward. On VideoNet-Gen, ERA reaches 0.50 under pressure against 0.53 with real returns, whereas SFT-Mix falls from 0.49 to 0.31 and the conventional RL rewards fall to 0.33–0.36; the ordering is the same on CG-Bench and NExT-GQA. The hallucination rate shows the same gap: SFT-Mix and Judge-SFT overrely on 0.41 and 0.47 of the swapped trajectories and the conventional RL rewards on 0.35–0.44, against 0.12 for ERA. ERA-RL has the lowest rate (0.05) but answers fewer questions (Table 4). Evaluation adds up to three assertive messages while training uses one, so this robustness extends beyond the training condition.

Table 4: **Re-seek behavior** on VideoNet-Gen. N: plain messages; P: assertive messages. RS-U: share of swapped trajectories that re-seek after the uninformative return; RS-D: share of trajectories that re-seek after a decisive first return; Hal.: overrely share; Ans.: share of trajectories that answer within $T_{\max}$ turns. Small numbers give the change from the row's start point (Qwen3.5-9B or SFT-Mix).

Model	Acc. ↑	RS-U N ↑	RS-U P ↑	RS-D P	Hal. P ↓	Ans. P ↑
Qwen3.5-9B	0.49	0.00	0.00	0.00	0.02	0.99
SFT-Single	0.47 ↓.02	0.05 ↑.05	0.01 ↑.01	0.01 ↑.01	0.43 ↑.41	0.99 =
SFT-CF	0.48 ↓.01	0.39 ↑.39	0.04 ↑.04	0.05 ↑.05	0.45 ↑.43	0.98 ↓.01
SFT-Mix	0.49 =	0.39 ↑.39	0.04 ↑.04	0.06 ↑.06	0.41 ↑.39	0.97 ↓.02
+ Outcome (per-call cost)	0.52 ↑.03	0.49 ↑.10	0.21 ↑.17	0.18 ↑.12	0.38 ↓.03	0.96 ↓.01
+ Outcome (redundant-call cost)	0.52 ↑.03	0.33 ↓.06	0.20 ↑.16	0.16 ↑.10	0.35 ↓.06	0.95 ↓.02
+ MM-Judge	0.50 ↑.01	0.37 ↓.02	0.12 ↑.08	0.10 ↑.04	0.44 ↑.03	0.97 =
+ **ERA**	0.53 ↑.04	0.63 ↑.24	0.40 ↑.36	0.14 ↑.08	0.12 ↓.29	0.93 ↓.04
ERA-RL (no SFT)	0.44 ↓.05	0.21 ↑.21	0.33 ↑.33	0.10 ↑.10	0.05 ↑.03	0.62 ↓.37

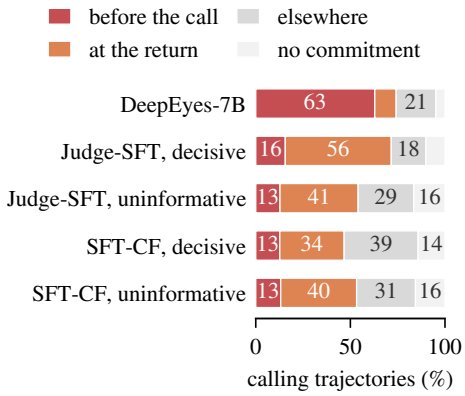

Figure 4: **Where the commitment point falls** in trajectories that call. DeepEyes-7B commits before the call; Judge-SFT commits at the return whether it is decisive or uninformative; SFT-CF commits at an uninformative return about as often as at a decisive one.

ERA on an existing agent. The ERA reward also improves an agent from prior work. On VideoNet-Gen, VideoZoomer+ERA keeps its accuracy with real returns (0.36 to 0.37), raises its accuracy under pressure from 0.22 to 0.37, and lowers its hallucination rate from 0.55 to 0.14, with similar gains on CG-Bench and NExT-GQA.

4.3 Analysis and Ablations

SFT and RL. The two SFT sets reach similar accuracy (0.47 and 0.48; Table 4), but only SFT-CF learns to re-seek, after 0.39 of the uninformative returns under plain messages against 0.05 for SFT-Single. Under pressure, this rate falls to 0.04, and 0.45 of its swapped trajectories overrely. Starting from SFT-Mix, every RL reward raises re-seek under pressure, but only ERA makes it depend on the return: ERA re-seeks after 0.40 of the uninformative first returns and 0.14 of the decisive ones, whereas the conventional rewards re-seek at similar rates after both (0.12–0.21 and 0.10–0.18). Without SFT, ERA-RL still learns to re-seek (0.33 under pressure) and rarely overrelies (0.05), but it answers only 0.62 of the questions because many trajectories keep re-seeking until the turn limit. These results indicate that SFT supplies the re-seek behavior, and the ERA reward ties it to what the returned frames show.

Failure modes of existing recipes. DeepEyes-7B, trained with tool-calling rewards, overlooks the return: 63% of its trajectories commit before the call (Figure 4), and continuing 62 trajectories from the first call with the real return, a random segment, a gray image, or $E^\star$ leaves the answer unchanged for 97% of the questions. Replacing its pre-call reasoning with a sentence stating that the coarse frames are insufficient restores the dependence on the return, with accuracy 0.45 under $E^\star$ against 0.18 under a random segment (p=0.031). Judge-SFT, fine-tuned on trajectories that Gemini judged to use the returned frames, overrelies: its commitment point moves to the return (56% of calling trajectories), but its post-return high-entropy token rate barely depends on the return (0.61

Table 5: **Ablations of ERA** on VideoNet-Gen, with the columns of Table 4; calls/q: calls per question; over.: overlook share. Small numbers give the change from the full ERA. †: one seed.

Variant	RS-U N↑	RS-U P↑	calls/q	over.↓	Hal. P↓	Acc. P↑
ERA	0.63	0.40	0.83	0.52	0.12	0.50
w/o location term†	0.20 ↓.43	0.13 ↓.27	0.48 ↓.35	0.77 ↑.25	0.24 ↑.12	0.41 ↓.09
w/o contrastive term†	0.12 ↓.51	0.06 ↓.34	0.67 ↓.16	0.61 ↑.09	0.36 ↑.24	0.38 ↓.12
binary location term	0.45 ↓.18	0.22 ↓.18	0.50 ↓.33	0.77 ↑.25	0.27 ↑.15	0.42 ↓.08
trajectory-level credit†	0.56 ↓.07	0.35 ↓.05	1.39 ↑.56	0.30 ↓.22	0.18 ↑.06	0.44 ↓.06
per-call cost†	0.55 ↓.08	0.25 ↓.15	0.17 ↓.66	0.93 ↑.41	0.21 ↑.09	0.40 ↓.10
gate from the start	0.30 ↓.33	0.16 ↓.24	0.55 ↓.28	0.74 ↑.22	0.20 ↑.08	0.45 ↓.05
gate after the first epoch	0.51 ↓.12	0.29 ↓.11	0.74 ↓.09	0.63 ↑.11	0.15 ↑.03	0.47 ↓.03

and 0.59), and it re-seeks on 0.03 of the questions. VideoZoomer re-seeks regardless of the return, after 0.95 of 385 uninformative and 0.96 of 27 decisive first returns on CG-Bench.

Ablations. Both process terms are needed for re-seek (Table 5). Without the contrastive term, the agent still calls but re-seeks after only 0.12 of the uninformative returns and overrelies on 0.36 of the swapped trajectories; without the location term, calls fall to 0.48 and the overlook share rises to 0.77, the same as with a binary location term. Crediting the process terms to the whole trajectory raises calls to 1.39 per question, with calls on 0.82 of the control questions against 0.42 under ERA, while charging every call cuts calls to 0.17 and raises the overlook share to 0.93. Gating correctness from the start or after the first epoch lowers re-seek to 0.30 and 0.51, whereas switching the gate on once the re-seek rate exceeds η gives the highest re-seek (0.63) and the lowest hallucination rate (0.12).

5 Conclusion

An agent uses a tool return only if its decision is still open when the return arrives. Reading this from the agent's high-entropy tokens, ERA exposes the overlook and overrely failures of current recipes and, as an RL reward, ties the re-seek that SFT teaches to what the returned frames show, even under assertive messages.

AI Use Statement

We used generative AI tools to assist with writing, namely to edit the paper for readability and to suggest its structure, and, as described in Section 4.1, Gemini generated the multiple-choice questions of VideoNet-Gen and the re-seek demonstrations of SFT-CF. The authors reviewed all AI-assisted text and take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

Ethics Statement

This work uses public video benchmarks and involves no human subjects. VideoNet-Gen questions were generated by a proprietary model and filtered manually, and they will be released only alongside the corresponding public VideoNet videos under their original licenses.

Reproducibility Statement

Section 3 defines the entropy trace, the commitment point, and the reward, and Section 4.1 and Table 2 give the data, benchmarks, and hyperparameters. Code and the trained model (*era-zoomer*) are released anonymously on Hugging Face, and the supplementary material contains anonymized code for the training environment, the reward, and the certification of tool-required questions, together with the prompts and the lists of certified questions.

REFERENCES

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, et al. CG-Bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, et al. Scaling RL to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, et al. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 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.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, et al. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, et al. Video-MME: The First-Ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Zhenkun Gao, Yicheng Bao, Jinlong Peng, Xueheng Li, Theo Huang, Bangwei Liu, et al. VideoSearcher: Empowering video deep research with Multi-Tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Jiawei Guo, Donglei Yu, Yu Chen, Xiang Wang, Shuai Li, Xinpei Zhao, et al. Do multimodal agents really benefit from tool use? a systematic study of capability gains. *arXiv preprint arXiv:2606.02357*, 2026.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. FrameThinker: Learning to think with long videos via Multi-Turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025.

David V. Hinkley and Elizabeth A. Hinkley. Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488, 1970. doi: 10.1093/biomet/57.3.477.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. DeepEyesV2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Haojie Huang, Xinlei Yu, Chengming Xu, Zhangquan Chen, Cheng Yang, Qingdong He, et al. Evidence-RL: Towards evidence-intensive visual reasoning. *arXiv preprint arXiv:2608.08021*, 2026.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, et al. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*, 2025.

Senjie Jin, Peixin Wang, Boyang Liu, Xiaoran Fan, Shuo Li, Zhiheng Xi, et al. Entropy is not enough: Unlocking effective reinforcement learning for visual reasoning via Vision-Anchored token selection. *arXiv preprint arXiv:2606.03937*, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, et al. Measuring faithfulness in Chain-of-Thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Wenqi Liu, Shijie Ma, Yunxiao Wang, Meng Liu, Qile Su, Han Liu, et al. Thinking beyond videos: Unifying video reasoning and deep research for Open-World video agents. *arXiv preprint arXiv:2608.23329*, 2026a.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, et al. VideoTemp-o3: Harmonizing temporal grounding and video understanding in agentic Thinking-with-Videos. *arXiv preprint arXiv:2602.07801*, 2026b.

Zhengyuan Pan, Yanhao Chen, Zhongquan Jian, Wanru Zhao, Haonan Ma, Meihong Wang, and Qingqiang Wu. Prototype entropy alignment: Reinforcing structured uncertainty in LLM reasoning. In *AAAI*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026. Official citation from the Qwen3.5 Hugging Face model card; no arXiv technical report available.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, et al. TSPO: Temporal sampling policy optimization for long-form video language understanding. *arXiv preprint arXiv:2508.04369*, 2025.

Giorgio Taricco. Entropy regularization in deep reinforcement learning: A structured review across classical control, generative policies, and reasoning language models. *Entropy*, 28(7):811, 2026. doi: 10.3390/e28070811.

Changyuan Tian, Zhicong Lu, Huaxing Liu, Xiang Wang, Shuai Li, Yu Chen, et al. Faithful-MR1: Faithful multimodal reasoning via anchoring and reinforcing visual attention. *arXiv preprint arXiv:2605.22072*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in Chain-of-Thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Haoqing Wang, Xingrun Xing, Wei Xia, Ziheng Li, and Yehui Tang. FaithEyes: Towards faithful tool use via Multi-Agent Process-Image verification. *arXiv preprint arXiv:2607.28225*, 2026a.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing Pixel-space reasoning with Curiosity-Driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, et al. Beyond the 80/20 rule: High-Entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025b.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, et al. Video-Thinker: Sparking "thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025c.

Zhiheng Wang, Bo Peng, Lai Wei, and Chaochao Lu. The illusion of visual Tool-Use: A causal audit of thinking with images. *arXiv preprint arXiv:2608.06270*, 2026b.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, et al. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025d.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, et al. MMSearch-R1: Incentivizing LMMs to search. *arXiv preprint arXiv:2506.20670*, 2025.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *CVPR*, 2024.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, et al. DeFacto: Counterfactual thinking with images for enforcing Evidence-Grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025.

Tanush Yadav, Mohammadreza Salehi, Jae Sung Park, Vivek Ramanujan, Hannaneh Hajishirzi, Yejin Choi, et al. VideoNet: A Large-Scale dataset for Domain-Specific action recognition. *arXiv preprint arXiv:2605.02834*, 2026. CVPR 2026. Dataset: `https://huggingface.co/datasets/raivn/VideoNet`.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, et al. LongVT: Incentivizing "thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An Open-Source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, et al. Frame-Voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, et al. VideoExplorer: Think with videos for agentic Long-Video understanding. *arXiv preprint arXiv:2506.10821*, 2025.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, et al. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *NeurIPS Datasets and Benchmarks*, 2023.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, et al. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

A Discussion and limitations

Position, not confidence. A confidence or answer-margin measure captures *how sure* the agent is at one moment; such a measure needs an answer format and saturates once training has made the agent confident. The entropy trace measures *when* the agent stopped being unsure, and its informative part is a position rather than a level. This property lets a single rule both diagnose an agent and credit a turn, which an LLM-as-a-Judge, seeing the same words on either side of an uninformative return, cannot do (Section 4.3).

Why video agents commit to an answer before the tool call returns. Both the overlook and overrely failures stem from their training signals: neither an outcome reward nor an LLM-as-a-Judge compares the same reasoning under two different returns, and such a comparison is necessary to define evidence use. The contrastive term makes this comparison explicit; without the contrastive term, the location term alone could be satisfied by committing late on every return.

Beyond video. Nothing in the entropy trace or the entropy-change reward refers to frames. Any tool whose return can be replaced by an uninformative return of the same form—an image crop, a retrieved passage, an execution result—admits the same construction.

Limitations. The commitment point is a behavioral statistic, not a measurement of an internal state, and we validate this statistic by intervention. The location term uses the evidence windows $E^\star$ during training; certification is relative to the certifying model. Our agents are 7–9B models trained with LoRA, and uninformative returns keep the requested time labels.

Can the reward be gamed? Three shortcuts are conceivable. (i) *Never commit*: a flat trace scores $0 > -1$. It forfeits the $+1$, and with at most four turns an agent that keeps searching past decisive evidence pays the redundant-call cost and loses the outcome term. (ii) *Always call twice*: in our environment a re-seek is not handed $E^\star$ (Section 3.4), so a re-seek helps only if the new window is better, and a re-seek after a decisive return is charged. (iii) *Hedging filler* to push $\tau^\star$ later or to inflate U_{ins}: the trace marks a fixed share of tokens per trajectory, and the contrastive term compares two continuations of the same prefix, so filler present under both returns cancels out. Policy gradient acts on sampled tokens rather than on entropy itself; the entropy-change reward changes the trace by selecting, among the G rollouts of a question, those whose high-entropy tokens fall where the terms encourage them.

B ADDITIONAL CASES

Figures 5 and 6 show two further test questions on which both agents answer incorrectly, so that accuracy cannot separate them either. In both, the outcome-trained agent commits during its first turn. In Figure 5 it reads the requested step into frames that show a later moment, and in Figure 6 it re-seeks but its commitment point lies before the first call, so the second return changes nothing. The ERA agent states in each case what the uninformative return lacks, requests the moment it identified from the coarse frames, reaches the evidence window $E^\star$, and commits only afterwards. Its final answers are wrong because the decisive frames are misread, not because the evidence was unseen. The quoted reasoning is translated from the agent's Chinese output; the returned frame shown is the middle frame of each return.

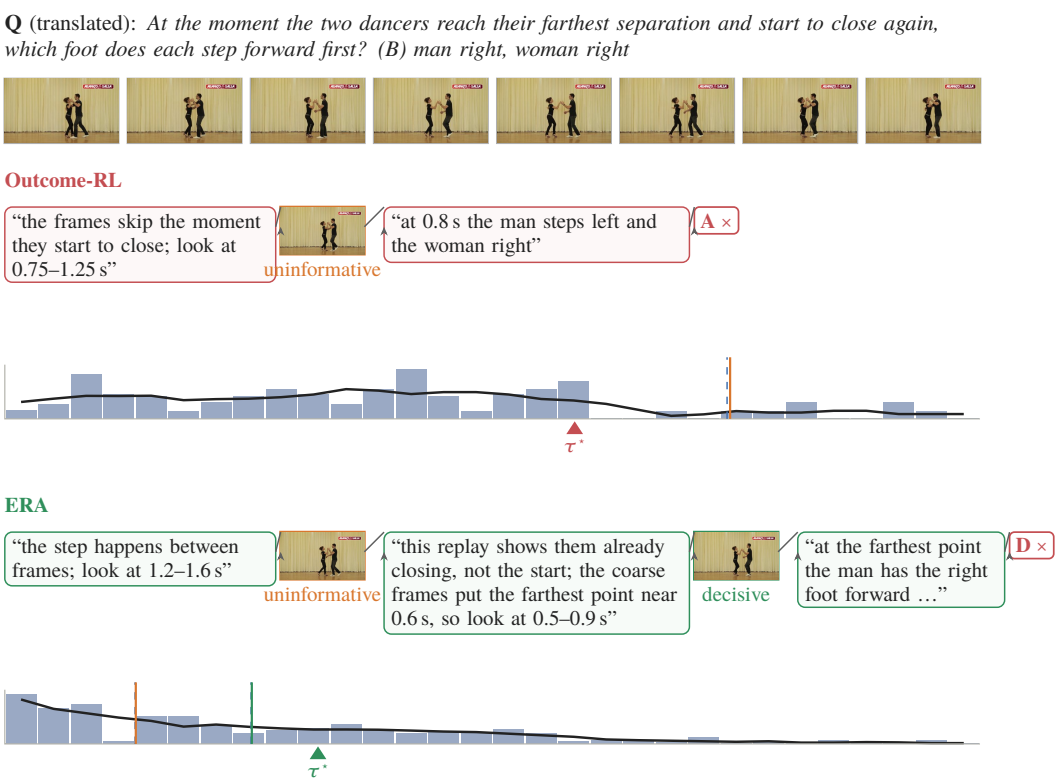

Figure 5: **Dance case.** The outcome-trained agent commits before its call (overlook) and answers from an uninformative return; the ERA agent notices that the return shows the dancers already closing, re-seeks the farthest point and commits after the decisive return (evidence-resolved).

Q (translated): *At the instant the dunking player takes off, which foot leaves the ground last and how are his legs held in the air? (D) left foot last, right knee raised*



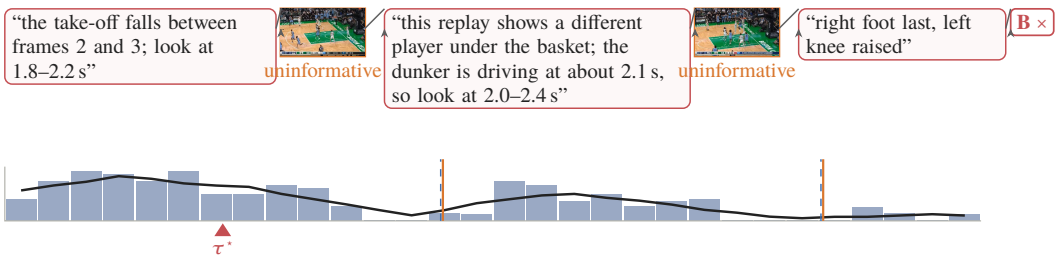

Figure 6: **Dunk case.** Both agents re-seek, but the outcome-trained agent's commitment point precedes its first call and its second request also misses the evidence window; the ERA agent identifies the dunker's drive in the coarse frames, requests 1.3–2.0 s and commits after the decisive return.