RELIABLE REASONING FOR VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal chain-of-thought (CoT) has emerged as a promising paradigm for improving the reasoning capabilities of multimodal large language models (MLLMs): It typically extracts question-relevant cues from a static image and reasons over them to derive an answer. The success of this paradigm partly rests on the assumption that the static image contains sufficient evidence for the question. Videos, however, challenge this assumption because they are temporally structured sequences of frames in which relevant evidence can be distributed unpredictably over time. This introduces two sources of unreliability. First, a single sampled video observation may omit crucial evidence, causing the model to reason from an under-informed view; we refer to this as unreliability arising from under-informed reasoning. Second, the model's understanding and answer may change substantially as richer observations become available; we refer to this as unreliability arising from cross-observation inconsistency. To address both challenges, we introduce Reliable Reasoning with Videos (RRV), which assesses answer reliability at two levels. Within a single observation, RRV uses video information extraction and answer reasoning spans when both are available, falling back to whole-response uncertainty otherwise, and aggregates the reliability-weighted trajectories into an answer distribution. Across observations, RRV checks answer consistency along a nested sequence of increasingly informative frame sets until two adjacent observations propose the same answer. By estimating the reliability of CoT within each observation and filtering answers based on consistency across observations, RRV suppresses the influence of unreliable CoT reasoning on final predictions. Experiments on three video MLLMs and eight video QA benchmarks show that RRV improves accuracy in most evaluated settings.

1 INTRODUCTION

Chain-of-thought (CoT) reasoning improves the ability of large language models to solve complex reasoning tasks by generating explicit intermediate steps (Wei et al., 2022; Kojima et al., 2022). Multimodal CoT incorporates visual information into this process (Zhang et al., 2024), extracting question-relevant cues from an image and reasoning over them to derive an answer (Zheng et al., 2023a; Li et al., 2026). Answer quality therefore depends on accurate visual understanding and sound reasoning from the available evidence.

Existing methods improve answer reliability by using answer agreement, model confidence, or token-level uncertainty to aggregate, weight, or filter trajectories (Wang et al., 2023; Kadavath et al., 2022; Taubenfeld et al., 2025; Kang et al., 2025; Fu et al., 2025b; Li et al., 2026). Adaptive sampling reduces cost by stopping when sampled answers are sufficiently consistent (Aggarwal et al., 2023; Li et al., 2024c). When the visual input remains fixed during generation and aggregation, however, these reliability estimates are conditional on the evidence in that input. Repeated reasoning can measure answer agreement and generation uncertainty under the available evidence, but cannot determine whether the answer will remain supported when additional visual evidence is introduced.

This limitation is particularly important in video reasoning. Models typically process sampled or compressed video frames (Li et al., 2024b; He et al., 2024; Song et al., 2024; Shen et al., 2025), while question-relevant evidence may appear briefly or at widely separated times (Buch et al., 2025; Wu et al., 2024). A single observation can thus omit key information, leading to under-informed reasoning. Trajectories generated from the same observation may share this omission and agree on an answer without sufficient evidence. Adding new frames can then change the answer, leading to

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

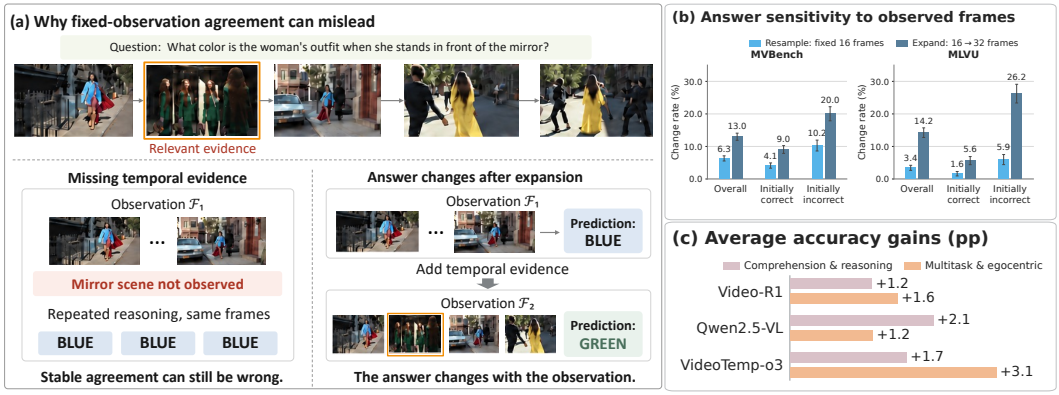


Figure 1: (a) Missing visual evidence can yield consistent but incorrect answers. (b) On MVBench and MLVU, Video-R1 changes answers more often after adding frames than after resampling reasoning trajectories from fixed frames. (c) Average accuracy gains of RRV over SC.

cross-observation inconsistency. Figure 1(a) illustrates this connection: the initial observation omits the scene in which the woman stands in front of a mirror, and repeated reasoning consistently answers "blue"; after frames containing the relevant evidence are added, the answer becomes "green." Agreement under a fixed observation does not imply support from more complete video evidence.

Figure 1(b) compares answer changes under repeated generation from fixed frames and under frame expansion. With the same number of voting trajectories and the same initial answer as the reference, Video-R1 (Feng et al., 2025) changes answers more often after expansion on MVBench and MLVU, with a larger difference for initially incorrect answers. Error bars show 95% paired video-cluster bootstrap confidence intervals. These results show that repeated reasoning under a fixed observation does not fully capture answer sensitivity to additional visual evidence. Reliable video reasoning requires assessing trajectory reliability within the current observation and checking whether the answer remains supported as the evidence expands.

To address these challenges, we introduce Reliable Reasoning with Videos (RRV), which combines within-observation reliability estimation with cross-observation stability verification. Within each observation, RRV weights trajectories using uncertainty in video information extraction and answer reasoning, scoring the complete response when these stages cannot be separated. When both stages are available, the less reliable stage determines the weight. Weighted answers form an observation-specific distribution, and sampling stops early once the remaining budget cannot change its leader. RRV then adds frames while retaining earlier ones and estimates a new answer distribution independently. It returns early only when two adjacent observations share the same stable leader; otherwise, it continues to the final observation. The lower support for the returned answer in the final adjacent pair provides its reliability ranking score.

Experiments compare RRV with existing methods on video question answering. Figure 1(c) summarizes mean accuracy gains over SC for comprehension and reasoning in Table 1, and multitask and egocentric understanding in Table 2, averaging benchmarks equally within each group. Gains range from 1.2 to 2.1 and from 1.2 to 3.1 percentage points, respectively. These experiments complement our theoretical analysis of the limits of fixed-observation agreement and the role of trajectory weighting in answer selection.

Our contributions are summarized as follows:

- We identify under-informed reasoning and cross-observation inconsistency as two distinct sources of unreliability in video reasoning, and show that agreement under a fixed observation does not guarantee answer stability as the evidence expands.

- We propose RRV, which combines uncertainty-weighted trajectory aggregation with adaptive sampling within each observation. It progressively expands the frame set and returns early only when adjacent observations share the same stable leader.

- Across three video MLLMs and eight benchmarks, RRV ranks first in 20 of 24 comparisons (19 outright and one tied) with the reported baselines at different inference budgets. Further experiments assess reliability, sampling choices, and accuracy versus inference cost.

2 Related Work

Video reasoning from partial observations. Video MLLMs align sampled frames with language representations for video understanding (Maaz et al., 2024; Lin et al., 2024; Jin et al., 2024). Long-video systems reduce visual context through compact tokenization, sparse memory, or adaptive compression (Li et al., 2024b; He et al., 2024; Song et al., 2024; Shen et al., 2025; Korbar et al., 2024), while temporally grounded models and benchmarks emphasize retrieval and reasoning over localized events in long videos (Ren et al., 2024; Huang et al., 2024; Qian et al., 2024; Fu et al., 2025a; Wu et al., 2024). Building on adaptive frame recognition (Wu et al., 2019; Ghodrati et al., 2021), recent video MLLMs select, retrieve, or progressively acquire question-relevant frames (Hu et al., 2025; Wang et al., 2024; 2025; Liao et al., 2024; Buch et al., 2025; Yao et al., 2025; He et al., 2026; Xie et al., 2026). These methods improve which visual evidence reaches the model, but do not assess whether a prediction remains stable after previously unseen temporal evidence is added.

Reliability of chain-of-thought reasoning. Self-consistency improves reasoning by sampling multiple trajectories and marginalizing their answers (Wang et al., 2023). Adaptive variants stop once the answer distribution stabilizes (Aggarwal et al., 2023; Li et al., 2024c), while test-time scaling studies how to allocate additional inference compute (Snell et al., 2025). Other methods reweight or filter trajectories using output confidence or token-level uncertainty (Taubenfeld et al., 2025; Kang et al., 2025; Fu et al., 2025b; Li et al., 2026). Calibration, selective prediction, and model self-evaluation instead assess reliability through probabilities, generation disagreement, or semantic entropy (Jiang et al., 2021; Tian et al., 2023; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019; Yoshikawa & Okazaki, 2023; Kadavath et al., 2022; Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Cole et al., 2023; Arana et al., 2026). Robustness to answer presentation has also been studied through option priors and permutation consistency (Zheng et al., 2024; Tang et al., 2024). These approaches vary decoding or answer presentation while holding the observed video fixed. RRV first stabilizes the answer within each observation using stage-wise trajectory reliability, then compares the stabilized answers across nested observations without pooling distributions conditioned on different frame sets.

3 Method

As illustrated in Figure 2, RRV assesses trajectory reliability within each observation and verifies answer stability as visual evidence expands. Within-observation reliability estimation combines balanced sampling over input presentations, reliability-weighted aggregation, and adaptive trajectory stopping. Cross-observation stability verification expands the observed frame set and checks adjacent answers. Section 3.1 introduces video CoT and SC. Section 3.2 defines within-observation reliability. Section 3.3 describes cross-observation stability. Section 3.4 specifies the stopping rules and reliability score. Section 3.5 presents the theoretical analysis.

3.1 Preliminaries

Video chain-of-thought reasoning. We consider video CoT models that generate intermediate textual reasoning from sampled visual observations (Feng et al., 2025; Liu et al., 2026). Let $\mathbf{V} = (v_1, \dots, v_T)$ be a video of T frames. A sampling rule π selects an ordered observation of b frames:

$$\mathcal{F}_{\pi,b}(\mathbf{V}) = (v_{t_1}, \dots, v_{t_b}), \qquad 1 \le t_1 < \cdots < t_b \le T. \tag{1}$$

Together with question $\mathbf{q}$, these frames form input $x_{\pi,b} = (\mathcal{F}_{\pi,b}(\mathbf{V}), \mathbf{q})$. Indexing sampled trajectories by i, the model generates $\mathbf{t}_i = (z_{i,1}, \dots, z_{i,L_i})$, where u indexes its generated tokens, and extracts an answer as follows:

$$p_\Theta(\mathbf{t}_i \mid x_{\pi,b}) = \prod_{u=1}^{L_i} p_\Theta(z_{i,u} \mid z_{i,<u}, x_{\pi,b}), \qquad a_i = g_{\mathbf{q},x_{\pi,b}}(\mathbf{t}_i) \in \mathcal{O}_\mathbf{q} \cup \{\perp\}. \tag{2}$$

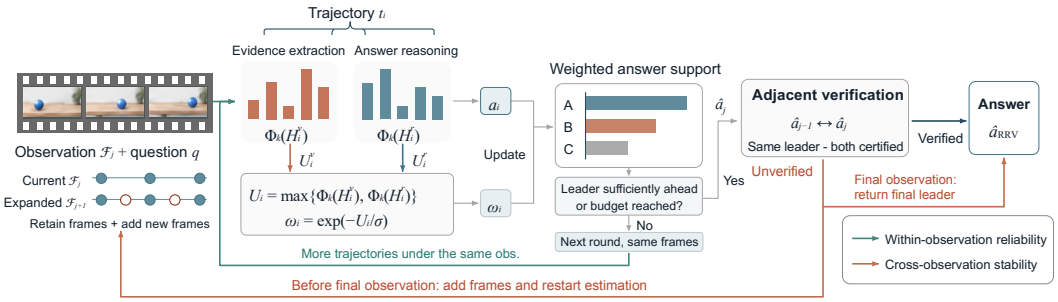


Figure 2: Overview of RRV, which aggregates answers using trajectory reliability within each observation and checks answer stability as visual evidence expands.

Here, $\perp$ denotes failure to extract a canonical answer. The fixed protocol $g_{\mathbf{q},x}$ converts a response to the canonical answer space $\mathcal{O}_{\mathbf{q}}$. For multiple-choice QA, it maps a displayed label to its option identity. For open-ended QA, one frozen semantic-grouping protocol jointly partitions all parseable candidate responses for a question and maps each response to its group.

Self-consistency. Given one fixed input x, self-consistency (SC) samples N trajectories and returns their modal answer (Wang et al., 2023):

$$p_{\text{SC}}(o \mid x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[a_i = o], \qquad \hat{a}_{\text{SC}} = \arg\max_{o \in \mathcal{O}_{\mathbf{q}}} p_{\text{SC}}(o \mid x). \tag{3}$$

Throughout, zero support for every canonical answer gives $\hat{a} = \perp$; other ties follow a fixed canonical order. SC estimates generation variation conditioned on the sampled frames. It does not test whether the answer changes when the video observation changes.

3.2 Within-Observation Reliability Estimation

Accordingly, RRV first forms an answer distribution for the j-th video observation $\mathcal{F}_j$ in the nested sequence introduced in Section 3.3, while accounting for differences in trajectory reliability. Within a fixed observation, multiple-choice QA marginalizes known sensitivity to option presentation (Zheng et al., 2024; Tang et al., 2024) using K predetermined, answer-preserving option orders, where K is at most the number of distinct orders a question admits; open-ended QA has one presentation. We denote the resulting inputs by $x_{j,1}, \ldots, x_{j,K}$, where $\ell \in \{1, \ldots, K\}$ indexes the presentation. Eq. (2) maps their predictions to the same answer space, and each sampling round draws one trajectory from every presentation.

Trajectory reliability. For a trajectory i sampled from $x_{j,\ell}$, we use next-token entropy to estimate trajectory uncertainty (Li et al., 2026). When the output format permits a split, a model-specific parser identifies two scoring spans, $\mathbf{t}_i^v$ and $\mathbf{t}_i^r$, as proxies for evidence extraction and answer reasoning. Their boundaries follow the available output structure; two semantically distinct stages are not assumed. Let H_i^v, H_i^r, and H_i denote the token-entropy sets of these spans and the complete response, respectively. We summarize any nonempty set H by the mean of its k largest values:

$$\Phi_k(H) = \frac{1}{\min\{k, |H|\}} \sum_{h \in \text{TopK}_{\min\{k,|H|\}}(H)} h. \tag{4}$$

For $H_i \neq \varnothing$, we score the more uncertain span when both exist, or the full response otherwise:

$$U_i = \begin{cases} \max\{\Phi_k(H_i^v), \Phi_k(H_i^r)\}, & H_i^v \neq \varnothing \ \wedge \ H_i^r \neq \varnothing, \\ \Phi_k(H_i), & H_i^v = \varnothing \ \vee \ H_i^r = \varnothing. \end{cases} \tag{5}$$

We then assign $\omega_i = \exp(-U_i/\sigma)$, where $\sigma > 0$ controls the effect of uncertainty. Section 4.4 examines the association between these weights and trajectory-answer correctness; Appendix A.1 specifies the scoring-span boundaries and entropy computation.

Answer support. These weights convert trajectory reliability into answer support within $\mathcal{F}_j$. Sampling proceeds in *balanced rounds*, each drawing one trajectory from every presentation, so that no presentation is over-represented. After r such rounds the observation-level distribution and its leading answer are:

$$p_j^{(r)}(o) = \frac{1}{K}\sum_{\ell=1}^{K} \frac{\sum_{i=1}^{r} \omega_{j,\ell,i}\mathbf{1}[a_{j,\ell,i} = o]}{\sum_{i=1}^{r} \omega_{j,\ell,i}}, \quad \hat{a}_j^{(r)} = \arg\max_{o \in \mathcal{O}_\mathbf{q}} p_j^{(r)}(o). \tag{6}$$

Eq. (6) uses trajectory reliability to form an answer distribution under $\mathcal{F}_j$, but all predictions share the same observed frames and therefore do not capture answer variation across frame sets. We discuss the effect of trajectory weighting on answer selection in Section 3.5, with a sufficient condition for preserving the unweighted leader given in Lemma 1.

3.3 Cross-Observation Stability Verification

To compare predictions for the same question across frame sets, RRV constructs $M \geq 2$ progressively expanded video observations. Let $[T] = \{1, \ldots, T\}$, let $I_j \subseteq [T]$ index the frames in observation $\mathcal{F}_j = (v_t)_{t \in I_j}$, and write $b_j = |I_j|$. The index sets satisfy:

$$I_1 \subsetneq I_2 \subsetneq \cdots \subsetneq I_M \subseteq [T], \qquad 2 \leq b_1 < b_2 < \cdots < b_M \leq T. \tag{7}$$

We initialize I_1 with uniformly spaced frames including both endpoints. Each I_{j+1} retains I_j and repeatedly adds one index until it holds b_{j+1} frames: among the gaps between consecutive members of the current set, take the widest, and add the index in $[T] \setminus I_{j+1}$ closest to its midpoint; ties in either choice favor earlier indices. Each step adds a new index, so the construction terminates whenever $b_M \leq T$. Each $\mathcal{F}_j$ follows Eq. (1) with $b = b_j$ along a fixed path to I_M. For retained answers $\hat{a}_j$, define adjacent agreement by:

$$C_j = \mathbf{1}[\hat{a}_{j-1} = \hat{a}_j \in \mathcal{O}_\mathbf{q}], \qquad j = 2, \ldots, M. \tag{8}$$

Here $C_j = 1$ indicates that a valid leader persists under one expansion. The indicator records answer agreement only; the return rule of Section 3.4 additionally requires both leaders to be certified.

3.4 Reliable Inference

Adaptive trajectory sampling. For each observation, RRV samples balanced rounds, drawing one trajectory per presentation in every round, up to $n_{\max} \geq 2$. With answer identities and existing weights fixed, $\underline{p}_{j,r}(o)$ and $\overline{p}_{j,r}(o)$ bound the support attainable at $n_{\max}$. We stop early only when a valid leader's lower bound exceeds every alternative's upper bound:

$$\underline{p}_{j,r}(\hat{a}_j^{(r)}) > \max_{o \neq \hat{a}_j^{(r)}} \overline{p}_{j,r}(o). \tag{9}$$

If the condition holds, the current leader is certified and retained as $\hat{a}_j$; otherwise, one more balanced round is sampled and the condition is checked again, until it first holds or $n_{\max}$ is reached. Appendix C.1 gives the bounds, missing-probability fallback, and tie handling; Proposition 1 shows that certification requires more than half the budget.

Cross-observation stopping. RRV traverses the nested observations from $\mathcal{F}_1$ and returns at the first $j \geq 2$ with $C_j = 1$ and both $\hat{a}_{j-1}$ and $\hat{a}_j$ certified, giving $\hat{a}_{\text{RRV}} = \hat{a}_j$; if no such j exists it reaches $\mathcal{F}_M$ and returns $\hat{a}_M$. The return observation $j^\star$ therefore satisfies $2 \leq j^\star \leq M$ in both cases.

Reliability score. Let r_j be the number of rounds used at observation j. For a valid returned answer, its score is the weaker support in the final adjacent pair $(\mathcal{F}_{j^\star-1}, \mathcal{F}_{j^\star})$:

$$c_{\text{RRV}} = \min\left\{ p_{j^\star-1}^{(r_{j^\star-1})}(\hat{a}_{\text{RRV}}), p_{j^\star}^{(r_{j^\star})}(\hat{a}_{\text{RRV}}) \right\}. \tag{10}$$

The score is unavailable when $\hat{a}_{\text{RRV}} = \perp$. Denote the preceding- and return-observation supports of the returned answer by s_- and s_+, respectively; thus $c_{\text{RRV}} = \min(s_-, s_+)$. At an early return, both terms score the same certified leader. A fallback at $\mathcal{F}_M$ can follow either adjacent disagreement or a failure to certify an agreeing pair; it does not necessarily imply low preceding support. This score summarizes adjacent support. It is neither a calibrated probability of correctness nor an additional stopping rule. Appendix A.6 evaluates its ranking and selective risk.

3.5 Theoretical Analysis

We now analyze the inference rule of Section 3.4 under the assumptions of Appendix C, which also contains the proofs and the cost analysis of the nested schedule. The results give a margin condition under which a leader survives a change of its support, characterize when the certificate fires, and bound the error of an early return.

Lemma 1 (margin stability). Let p and q be supports on $\mathcal{O}_{\mathbf{q}}$, $\hat{a}_q = \arg\max_o q(o)$, and m_q the gap between the two largest values of q. Under the same tie rule:

$$\max_o |p(o) - q(o)| < m_q/2 \implies \arg\max_o p(o) = \hat{a}_q. \tag{11}$$

With q the unit-weight and p the weighted support of the same trajectories, weighting cannot change the leader unless it moves some support by at least half the unweighted margin.

Proposition 1 (certificate). With sampled identities and weights fixed, once Eq. (9) holds after round r, $\hat{a}_j^{(r)}$ is the leader of every completion to $n_{\max}$ rounds; certification never changes the full-budget answer. Let $\bar{m}_{j,r} \in [0,1]$ be the current leader's presentation-averaged count margin over its strongest competitor. With unit weights, Eq. (9) holds after round r if and only if:

$$r\,(1 + \bar{m}_{j,r}) > n_{\max}. \tag{12}$$

Sampling therefore stops at the first round above $n_{\max}/(1+\bar{m}_{j,r})$, and under any weights never at $r \le n_{\max}/2$.

Proposition 2 (early return). Let $\hat{a}_j$ be the leader retained at $\mathcal{F}_j$, p_j its support at the retained round with margin m_j, and p_M the full-budget support at $\mathcal{F}_M$. Pathwise, $\hat{a}_j$ remains the leader of p_M unless $\max_o |p_M(o) - p_j(o)| \ge m_j/2$; in particular, a returned answer with $c_{\mathrm{RRV}} > 1/2$ is overturned only if its support falls by more than $c_{\mathrm{RRV}} - 1/2$. Let P_j be the population support at $\mathcal{F}_j$ (Appendix C.4) with leader $a_j^\star$ and margin $\Delta_j > 0$. Under unit weights and independent sampling given each $x_{j,\ell}$:

$$\Pr(\hat{a}_j \neq a_M^\star) \le \Pr\big(\max_o |P_M(o) - P_j(o)| \ge \Delta_j/2\big) + (|\mathcal{O}_{\mathbf{q}}| - 1)\, e^{-K n_{\max} \Delta_j^2/2}, \tag{13}$$

whether or not sampling stopped early. The first term does not depend on K, $n_{\max}$, or σ, and the second vanishes as $K n_{\max}$ grows: more trajectories from the fixed frames cannot reduce the reversal probability below the probability that expansion moves the support by half the margin, which only added frames affect. Adjacent agreement itself is rarely spurious: if $a_{j-1}^\star \neq a_j^\star$, then $\Pr(\hat{a}_{j-1} = \hat{a}_j)$ is at most the sum of the sampling terms at $\mathcal{F}_{j-1}$ and $\mathcal{F}_j$.

4 Experiments

We evaluate RRV in terms of answer accuracy, reliability, and inference cost, guided by the following four research questions:

- **RQ1.** Does RRV improve video QA across models and benchmarks?
- **RQ2.** How do progressive verification, weighting, and sampling budgets affect accuracy?
- **RQ3.** Can weights distinguish correct from incorrect answers across observation budgets?
- **RQ4.** How does RRV trade accuracy against inference cost?

4.1 Experimental Setup

Models and benchmarks. We evaluate three frozen video MLLMs: Video-R1-7B (Feng et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), and VideoTemp-o3 (Liu et al., 2026). The eight multiple-choice benchmarks are the long-video split of Video-MME (Fu et al., 2025a), LVB (Wu et al., 2024), VideoTemp-Bench (Liu et al., 2026), MMVU (Zhao et al., 2025), VSI-Bench (Yang et al., 2025), MVBench (Li et al., 2024a), MLVU (Zhou et al., 2025), and the labeled 500-question subset of EgoSchema (Mangalam et al., 2023).

Table 1: Video comprehension and reasoning accuracy (%).

Model	Method	Video-MME	LVB	VTB	MMVU	VSI-Bench	Avg.
Video-R1-7B	Pass@1	48.8	56.1	75.0	63.4	33.3	55.3
	SC	49.0	56.4	74.8	63.7	33.5	55.5
	Self-Cert	48.7	56.4	74.7	63.5	**34.0**	55.4
	CISC	49.2	56.5	74.8	64.0	33.5	55.6
	DeepConf	49.0	56.6	74.6	63.7	33.5	55.5
	RTWI	49.3	56.8	74.8	63.5	33.7	55.6
	RRV	**50.2**	**58.8**	**75.7**	**65.8**	33.0	**56.7**
Qwen2.5-VL-7B	Pass@1	44.1	51.4	70.1	54.2	35.9	51.1
	SC	48.9	56.6	73.8	61.1	35.3	55.2
	Self-Cert	45.4	52.1	70.5	54.2	35.9	51.7
	CISC	**49.8**	56.9	74.2	61.9	35.9	55.7
	DeepConf	49.4	56.8	73.3	60.8	35.1	55.1
	RTWI	49.6	57.7	73.5	61.0	34.7	55.3
	RRV	49.0	**59.4**	**75.8**	**65.1**	**37.1**	**57.3**
VideoTemp-o3	Pass@1	46.4	54.2	72.2	59.0	38.2	54.0
	SC	51.6	59.5	**76.7**	64.5	41.0	58.7
	Self-Cert	48.8	56.7	75.5	63.5	39.5	56.8
	CISC	51.6	59.2	76.6	65.6	40.7	58.7
	DeepConf	51.2	59.1	76.2	65.4	41.3	58.6
	RTWI	52.1	58.5	76.2	64.0	40.3	58.2
	RRV	**52.9**	**62.2**	76.1	**68.6**	**42.1**	**60.4**

Table 2: Video QA accuracy (%) on MVBench, MLVU, and EgoSchema.

Model	Dataset	Pass@1	SC	Self-Cert	CISC	DeepConf	RTWI	RRV
Video-R1-7B	MVBench	60.9	61.3	61.2	61.5	61.2	61.1	**63.4**
	MLVU	63.6	63.5	62.6	63.6	63.5	63.7	**63.8**
	EgoSchema	56.2	55.8	57.0	55.8	55.6	55.6	**58.2**
	Avg.	60.2	60.2	60.3	60.3	60.1	60.1	**61.8**
Qwen2.5-VL-7B	MVBench	57.1	61.2	57.9	61.5	61.2	60.9	**63.9**
	MLVU	56.6	**62.2**	57.7	**62.2**	60.8	61.5	61.8
	EgoSchema	62.4	68.6	62.4	67.8	68.2	67.2	**70.0**
	Avg.	58.7	64.0	59.3	63.8	63.4	63.2	**65.2**
VideoTemp-o3	MVBench	61.4	63.0	62.4	63.1	63.0	63.4	**66.4**
	MLVU	62.4	66.7	65.6	65.0	**66.9**	66.4	**66.9**
	EgoSchema	59.4	61.0	58.6	61.4	61.8	60.0	**66.6**
	Avg.	61.1	63.6	62.2	63.2	63.9	63.3	**66.6**

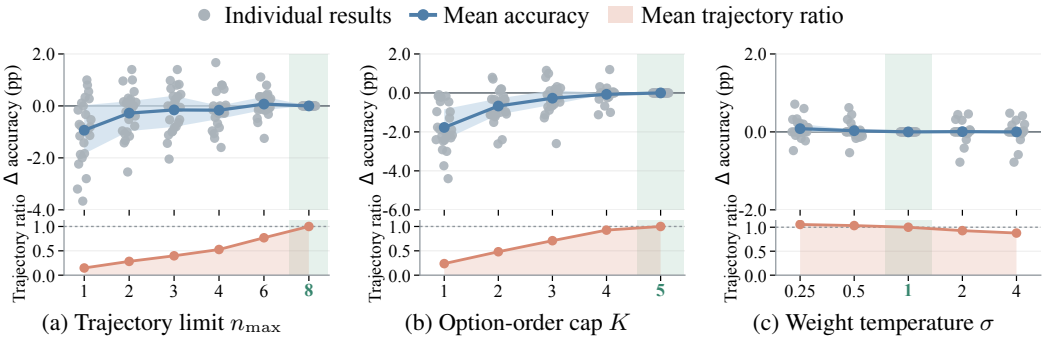
Baselines and metrics. We compare Pass@1, self-consistency (SC) (Wang et al., 2023), Self-Certainty (Kang et al., 2025), CISC (Taubenfeld et al., 2025), DeepConf (Fu et al., 2025b), and RTWI (Li et al., 2026). We report exact-match accuracy and inference cost. Accuracy and score metrics are percentages with one decimal place; accuracy differences are in percentage points. Within each backbone, methods share questions, 64 distinct source frames, preprocessing, prompts, and decoding settings. Inference uses vLLM (Kwon et al., 2023). Baselines use all 64 frames; RRV follows nested 16/32/64-frame observations with early stopping. RRV uses $K(q) = \min(5, d_q)$ cyclic option orders, including the canonical order, where d_q is the number of answer options. Trajectory and compute budgets differ. Appendix A provides implementation details.

4.2 Comparison with Existing Methods (RQ1)

Comprehension and reasoning. We first compare RRV with existing inference-time methods on five benchmarks covering video comprehension and reasoning. RRV ranks first in 12 of 15 comparisons in Table 1, with the highest average accuracy for every backbone. Average gains over standard

Table 3: Accuracy (%) of RRV under module ablations. Differences are computed before rounding.

Within-observation reliability estimation	Cross-observation stability verification	Video-R1		VideoTemp-o3		Qwen2.5-VL	
		MMVU	LVB	MMVU	LVB	MMVU	LVB
SC		63.7	56.4	64.5	59.5	61.1	56.6
×	×	63.4	56.1	59.0	54.2	54.2	51.4
✓	×	64.3	57.1	68.2	58.5	63.7	54.8
✓	✓	**65.8**	**58.8**	**68.6**	**62.2**	**65.1**	**59.4**



(a) Trajectory limit $n_{\max}$ (b) Option-order cap K (c) Weight temperature σ

Figure 3: Sensitivity of RRV to $n_{\max}$, K, and σ.

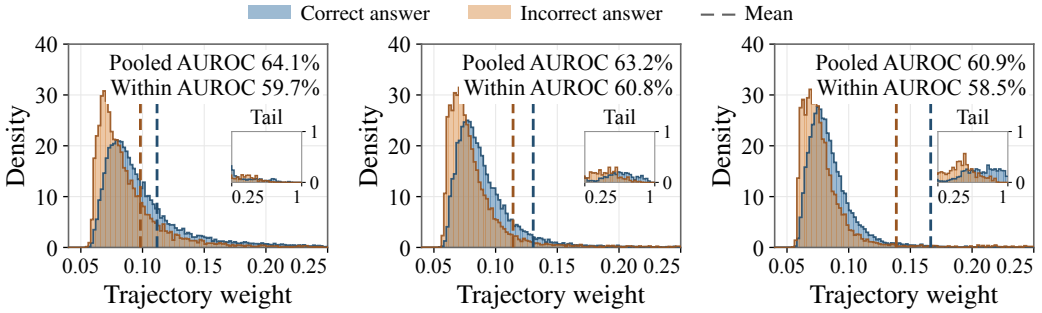
SC are 1.2, 2.1, and 1.7 percentage points for Video-R1, Qwen2.5-VL, and VideoTemp-o3, respectively. On MMVU, RRV exceeds CISC, the strongest baseline for each backbone, by 1.8 to 3.2 percentage points. It also leads on LVB for all three backbones. On the other three benchmarks, the best method varies with the backbone. For Qwen2.5-VL, the much larger gain over SC on MMVU than on Video-MME also shows that benefits vary substantially within one backbone.

Multitask and egocentric understanding. We next examine whether the gains extend to multitask and egocentric video understanding on MVBench, MLVU, and EgoSchema. RRV achieves the highest mean accuracy across these three benchmarks for every backbone. Across the 24 comparisons in both tables, RRV ranks first in 19 and ties for first in one at the reported precision. Margins over the strongest baseline are larger on MVBench and EgoSchema than on MLVU, where RRV shows a small lead, a tie, or a loss depending on the backbone.

4.3 Inference Design and Sensitivity (RQ2)

Module ablation. Table 3 compares single-trajectory inference, the within-observation module alone, and full RRV. The within-observation module improves over single-trajectory inference in all six comparisons, but falls below SC for VideoTemp-o3 and Qwen2.5-VL on LVB. Full RRV achieves the highest accuracy in all six comparisons. Adding cross-observation verification improves MMVU accuracy by 1.4, 0.5, and 1.4 percentage points for Video-R1, VideoTemp-o3, and Qwen2.5-VL, respectively; on LVB, the gains are 1.6, 3.7, and 4.6 percentage points. The larger improvements on LVB show that the relative contributions of the two modules vary across benchmarks. Further analyses appear in Appendix A.5.

Sampling and weighting. Figure 3 examines sampling and weighting across all 24 model and benchmark combinations. In full RRV, reducing $n_{\max}$ from eight to four saves 47.3% of trajectories on average, with a mean accuracy loss of 0.2 percentage points. Restricting RRV to one option order lowers mean accuracy by 1.8 percentage points while using 23.7% of the default trajectories on average. Each order retains a limit of eight trajectories, so changing K also changes the total sampling budget. The K sweep therefore measures the combined effects of option-order diversity and sampling allowance. Mean accuracy stays within 0.1 percentage points of the default across the tested temperatures from $\sigma = 0.25$ to 4, while the mean trajectory ratio decreases from 1.06 to 0.88. Shaded bands show interquartile ranges across the 24 comparisons. The sampling curves show diminishing average returns: increasing $n_{\max}$ from one to two recovers most of the mean accuracy



(a) 16 frames (b) 32 frames (c) 64 frames

Figure 4: Trajectory-weight distributions by answer correctness for VideoTemp-o3 on MMVU.

loss, whereas larger budgets yield smaller and nonmonotonic changes. The spread of individual results at small budgets also shows that a modest average loss can coexist with larger losses for particular models and benchmarks.

4.4 Trajectory Reliability (RQ3)

Reliability of trajectory weights. We test whether weights distinguish correct from incorrect trajectories within an observation. Figure 4 shows higher mean weights for correct trajectories at all three frame budgets for VideoTemp-o3 on MMVU. Within-question, within-presentation comparisons yield mean AUROCs of 59.7%, 60.8%, and 58.5% at 16, 32, and 64 frames, above chance (50.0%). On average, this is roughly six favorable rankings per ten within-group correct–incorrect pairs, with ties counted as half. Weights thus provide a relative reliability signal, but the distributions overlap substantially and more frames do not consistently improve separation. A high weight alone does not establish correctness; its effect on the final answer depends on each candidate's aggregate support.

4.5 Inference Cost (RQ4)

Effect of adaptive stopping. Table 4 shows that, on MMVU, stopping reduces RRV's output tokens by 45.7%, 54.9%, and 53.8%, and median latency by 59.9%, 72.4%, and 69.0%, for Video-R1, Qwen2.5-VL, and VideoTemp-o3, respectively, relative to no stopping. Certification cuts remaining within-observation draws; cross-observation agreement can skip the final 64-frame observation.

Table 4: Accuracy and inference cost on MMVU.

Method	Video-R1			Qwen2.5-VL			VideoTemp-o3		
	Acc. %	Tokens/SC	Time (s)	Acc. %	Tokens/SC	Time (s)	Acc. %	Tokens/SC	Time (s)
SC	63.7	1.00	21.37	61.1	1.00	14.40	64.5	1.00	44.34
RRV w/o stop	64.3	15.06	157.04	63.7	15.00	123.50	68.2	15.58	244.41
RRV	65.8	8.18	62.93	65.1	6.76	34.03	68.6	7.20	75.84

5 Limitations and Conclusion

Limitations. The main evaluation covers three frozen video MLLMs with a fixed 16/32/64-frame schedule. Future work could test other model families and adaptive frame selection. Shared source frames do not imply equal inference budgets: RRV consumes more trajectories than SC. The present results therefore do not establish an advantage at equal compute.

Conclusion. RRV combines uncertainty-weighted aggregation within observations with verification across nested frame sets. With answer mappings and weights fixed, its local certificate preserves the full-budget leader; adjacent agreement checks persistence under one frame expansion. The results show benchmark-dependent accuracy effects and reduced trajectory use relative to visiting all observations. Weighting gains depend on the inference policy. Sampling stability, evidence sensitivity, and correctness remain distinct.

Reproducibility Statement

Experimental settings and implementation details are provided in Section 4 and Appendix A.

AI use statement

Generative AI tools were used to assist with writing and revising the manuscript. The authors carefully reviewed and verified all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *EMNLP*, 2023.

Jasper Meynard Arana, Kristine Ann M. Carandang, Ethan Robert Casin, Christian Alis, and Christopher Monterola. When models hesitate: Answer instability as a label-free uncertainty signal for LLMs. In *ACL SRW*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. arXiv, 2025.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *CVPR*, 2025.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *EMNLP*, 2023.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *JMLR*, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *NeurIPS*, 2025.

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 *CVPR*, 2025a.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. arXiv, 2025b.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *ICML*, 2019.

Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. FrameExit: Conditional early exiting for efficient video recognition. In *CVPR*, 2021.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, 2024.

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. In *ICLR*, 2026.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-LLM based video frame selection for efficient video understanding. In *CVPR*, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *CVPR*, 2024.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *TACL*, 2021.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. arXiv, 2022.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-N selection for large language models via self-certainty. In *NeurIPS*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *ECCV*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, 2023.

Haobin Li, Yutong Yang, Yijie Lin, Xiang Dai, Mouxing Yang, and Xi Peng. Reliable thinking with images. arXiv, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024a.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *ECCV*, 2024b.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *ICLR*, 2024c.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. VideoINSTA: Zero-shot long video understanding via informative spatial-temporal reasoning with LLMs. In *EMNLP Findings*, 2024.

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 *EMNLP*, 2024.

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. In *ICML*, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *ACL*, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *ICML*, 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 *CVPR*, 2024.

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 *ICML*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *ICLR*, 2025.

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 *CVPR*, 2024.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *NAACL*, 2024.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In *ACL Findings*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*, 2023.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *ECCV*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

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 *CVPR*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, 2024.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. AdaFrame: Adaptive frame selection for fast video recognition. In *CVPR*, 2019.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced multi-turn reasoning for long video understanding. In *ICML*, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Boyang Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. LongVT: Incentivizing "thinking with long videos" via native tool calling. In *CVPR*, 2026.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *ACL Findings*, 2025.

Hiyori Yoshikawa and Naoaki Okazaki. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *EACL Findings*, 2023.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *TMLR*, 2024.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring expert-level multi-discipline video understanding. In *CVPR*, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023b.

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 *CVPR*, 2025.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Method **3**

3.1 Preliminaries . . . 3

3.2 Within-Observation Reliability Estimation . . . 4

3.3 Cross-Observation Stability Verification . . . 5

3.4 Reliable Inference . . . 5

3.5 Theoretical Analysis . . . 6

4 Experiments **6**

4.1 Experimental Setup . . . 6

4.2 Comparison with Existing Methods (RQ1) . . . 7

4.3 Inference Design and Sensitivity (RQ2) . . . 8

4.4 Trajectory Reliability (RQ3) . . . 9

4.5 Inference Cost (RQ4) . . . 9

5 Limitations and Conclusion **9**

A Supplementary Experiments **15**

A.1 Evaluation Settings . . . 15

A.2 Open-ended Video QA . . . 15

A.3 Accuracy and Cost . . . 16

A.4 Hallucination and Reasoning . . . 16

A.5 Additional Experimental Results . . . 17

A.6 Returned-Answer Reliability . . . 19

B Qualitative Analysis **20**

C Theoretical Details **21**

C.1 Certification (Proposition 1) . . . 21

C.2 Cost of the Nested Schedule . . . 21

C.3 Margin Stability (Lemma 1) . . . 22

C.4 Early Return (Proposition 2) . . . 22

A Supplementary Experiments

A.1 Evaluation Settings

Evaluation setup. Methods share questions, 64 distinct source frames, preprocessing, and decoding within each backbone. RRV uses $k = 20$, $\sigma = 1$, and both stopping rules. Generation uses temperature 1.0, top-p 0.95, and seed 2027. Within each cached analysis, identical presentations share trajectories; VideoTemp-o3 tools access only the current observation.

Data selection. We use the Video-MME long split, MMVU validation, VSI-Bench debiased test, and labeled EgoSchema subset. Numerical VSI-Bench questions and open-ended MMVU questions are excluded. MVBench retains 3,600 questions with at least 64 source frames and uses whole videos without question-specific temporal windows.

Baselines. In the main comparisons, baselines use one canonical presentation at 64 frames. Pass@1 takes the first trajectory; the others use eight. SC uses majority voting. Self-Certainty selects the valid trajectory with the highest mean full-vocabulary $\mathrm{KL}(U\|p)$, without Borda voting. CISC weights valid answers by a temperature-1 softmax of geometric-mean sampled-token probabilities. DeepConf averages negative log probabilities of the top-20 tokens at each position within 2,048-token windows (the whole response if shorter), uses the minimum window score, and filters below the valid trajectories' 10th confidence percentile before weighted voting. RTWI uses renormalized top-10 entropy, adaptive k, API-turn stages, stage-wise 40th-percentile filtering, and $\gamma = 1$; empty selections fall back to SC.

Stage parsing and reliability scores. We construct scoring spans from recorded assistant outputs. For Video-R1, the token that first completes <answer> in the cumulative decoded text starts $\mathbf{t}_i^r$; preceding tokens form $\mathbf{t}_i^v$. VideoTemp-o3 applies the same split to responses without tool calls when a tagged answer is present and both spans are nonempty. For trajectories with tool calls, we concatenate all earlier nonempty assistant spans into $\mathbf{t}_i^v$ and use the last nonempty assistant span as $\mathbf{t}_i^r$, excluding tool responses. If fewer than two nonempty scoring spans are available, we score the full response. RRV computes token entropy using natural logarithms and renormalized top-20 vocabulary probabilities. We average the 20 highest token entropies within each span, or all values for shorter spans, and apply Eq. (5). If token probabilities are unavailable, all trajectories in the affected presentation receive unit weights.

Multiple-choice evaluation. Final answers are extracted using backbone-specific parsers and mapped to canonical options for evaluation. We use exact matching; ties follow canonical order and invalid answers count as incorrect.

Scoring and parsing configuration. For all three backbones, reliability scoring uses the generated reasoning and final answer, following the stage-based scoring and whole-response fallback rules defined above. Final answers are extracted according to each backbone's output format, using tagged-answer parsing or normally terminated, nonempty plain-text responses; full-response extraction is retained where supported.

Open-ended evaluation. For open-ended QA, Qwen2.5-VL-72B-Instruct-AWQ at temperature 0 groups candidates without references, then scores the lowest-ID representative of each frozen group against references (Zheng et al., 2023b). Grouping and scoring use separate fixed prompts; predictions are frozen before scoring. Stopping is replayed after grouping all cached candidates; equivalence to grouping only the available prefix is unverified. These results therefore evaluate cached-candidate aggregation, not a validated online prefix-only grouping procedure. Failed grouping and invalid outputs retain their questions and sampling costs. VideoReasonBench shares this judge across methods; baselines use 76 replacement trajectories across 42 questions, while RRV retains its original trajectories.

A.2 Open-ended Video QA

Open-ended accuracy. RRV achieves the highest average accuracy for every backbone in Table 5, exceeding SC in all nine comparisons and ranking first in seven. It leads on ActivityNet-QA (Yu et al., 2019) and NExT-QA (Xiao et al., 2021) for every backbone. On VideoSIAH (Yang et al., 2026), it leads with VideoTemp-o3 but trails the strongest baseline with Video-R1 and Qwen2.5-VL.

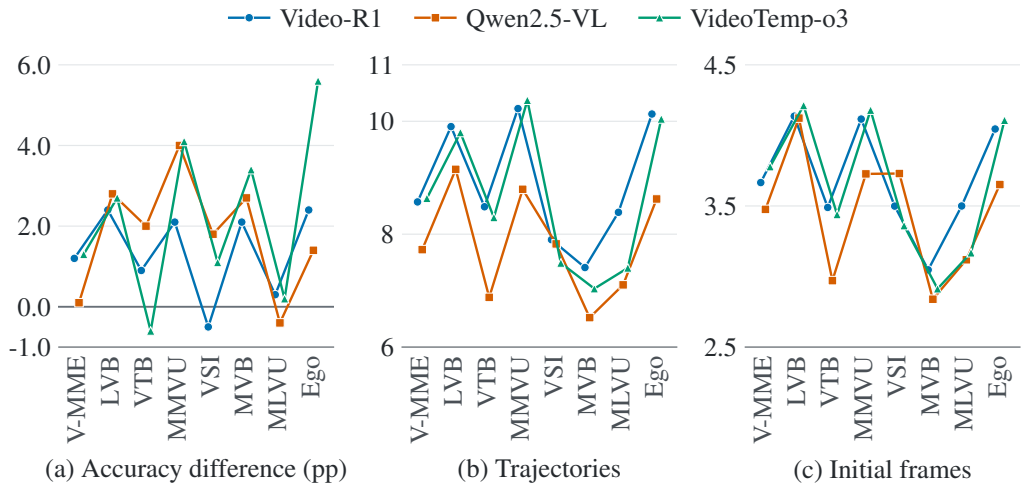
Within this cached-candidate protocol, Video-R1 gains most over SC on VideoSIAH, whereas Qwen2.5-VL and VideoTemp-o3 gain most on NExT-QA, showing that aggregation benefits depend on both backbone and benchmark.

Table 5: Open-ended video QA accuracy (%).

Model	Dataset	Pass@1	SC	Self-Cert	CISC	DeepConf	RTWI	RRV
Video-R1-7B	VideoSIAH	30.8	31.6	**33.0**	31.6	31.3	31.1	32.8
	ActivityNet-QA	51.9	52.0	51.7	52.1	52.1	51.9	**52.2**
	NExT-QA	43.1	44.1	43.5	43.9	43.9	43.7	**44.4**
	Avg.	41.9	42.6	42.7	42.5	42.4	42.2	**43.1**
Qwen2.5-VL-7B	VideoSIAH	24.4	30.5	27.1	**31.0**	30.1	**31.0**	30.7
	ActivityNet-QA	49.0	54.2	50.0	54.8	53.4	54.1	**55.0**
	NExT-QA	30.7	36.2	31.5	37.6	36.5	35.8	**38.1**
	Avg.	34.7	40.3	36.2	41.1	40.0	40.3	**41.3**
VideoTemp-o3	VideoSIAH	17.5	21.6	19.6	22.5	20.4	21.2	**22.9**
	ActivityNet-QA	37.8	45.7	40.1	45.7	45.4	43.8	**46.2**
	NExT-QA	34.1	38.6	34.8	38.6	38.2	37.5	**40.0**
	Avg.	29.8	35.3	31.5	35.6	34.7	34.2	**36.4**

A.3 Accuracy and Cost

Accuracy and cost. RRV exceeds SC in 21 of the 24 comparisons in Figure 5, with gains up to 5.6 percentage points and decreases no larger than 0.6 percentage points. Here the comparison is against SC; first-place counts in the main text compare all reported baselines. The separate resource measurements show higher trajectory and initial-frame consumption throughout. Initial-frame ratios are consistently smaller than trajectory ratios because many RRV trajectories use 16- or 32-frame observations, whereas SC always starts from 64 frames.



(a) Accuracy difference (pp) (b) Trajectories (c) Initial frames

Figure 5: Accuracy differences in percentage points (a) and separately measured trajectory and initial-frame cost ratios (b, c), relative to SC.

A.4 Hallucination and Reasoning

Hallucination and reasoning. Table 6 reports small accuracy differences for Video-R1. Relative to the strongest baseline, RRV answers nine more questions on EventHallusion, four more on VideoReasonBench, and one more on VCR-Bench. These point estimates do not establish a statistically

reliable advantage. On VideoHallucer, it exceeds SC but trails Self-Certainty. The low absolute accuracies on VideoReasonBench and VCR-Bench also show substantial remaining difficulty despite the gains from aggregation. VideoReasonBench also uses asymmetric trajectory replacements, as detailed in Appendix A; a matched replacement protocol is needed for a controlled comparison.

Table 6: Video-R1 accuracy (%) on hallucination and reasoning benchmarks.

	Binary yes/no		Open-ended	
Method	VideoHallucer	EventHall.	VideoReason.	VCR-Bench
n	2300	409	1440	524
Pass@1	70.5	48.2	5.9	18.5
SC	70.9	46.5	6.9	19.7
Self-Cert	71.7	45.5	6.5	19.3
CISC	71.0	46.7	6.9	19.7
DeepConf	70.7	47.2	7.0	19.7
RTWI	71.2	47.4	6.9	18.9
RRV	71.3	50.4	7.3	19.8

A.5 Additional Experimental Results

Weighting rules. Table 7 compares weighting rules on the same Video-R1/MMVU trajectories with cross-observation stopping enabled and local certification disabled. Accuracy varies by only 0.3 percentage points across the five rules. Mean ties with maximum for the highest accuracy and has the highest stratified AUROC, so the default maximum rule has no clear advantage here. Stratified AUROC averages within-group AUROC by group size, using each rule's own return groups; these groups differ across rules. Accuracy retains all questions, while AUROC excludes nonfinite scores.

Table 7: Accuracy and stratified AUROC of weighting rules on Video-R1/MMVU.

Metric	Uniform	Extraction	Reasoning	Mean	Maximum (RRV)
Acc. (%)	65.4	65.8	65.6	65.8	65.8
Stratified AUROC (%)	73.5	73.7	73.3	73.8	73.7

Stopping rules. Figure 6 compares four stopping policies. On the archived trajectories, enabling both rules reduces trajectory generation by 33.8% on average relative to no stopping. Certification preserves the full-budget leader within each observation and yields additional savings when combined with adjacent-agreement termination.

The MMVU and LVB accuracy differences computed from Table 3 are positive for all three backbones, ranging from 0.5 to 4.6 percentage points. The archived MLVU stopping comparisons show accuracy decreases of 1.7, 1.5, and 2.7 percentage points for Video-R1, Qwen2.5-VL, and VideoTemp-o3, respectively. Cross-observation termination can therefore improve or reduce accuracy, with the direction of the effect depending on the benchmark.

Diagnostic data sources. Figure 3 and Table 8 use the archived trajectory cache. In Figure 6, the six MMVU and LVB accuracy differences are computed from the full and within-observation accuracies in Table 3, which reports a separate replication under the same evaluation settings. All trajectory savings and the remaining accuracy differences use the archived stopping experiment. Accuracy differences are computed before rounding. In Table 4, no-stop accuracy follows the within-observation configuration in Table 3; tokens and latency are measured in the efficiency experiment.

Under the full RRV policy, the archived MLVU accuracies are 62.6%, 60.8%, and 65.0% for Video-R1, Qwen2.5-VL, and VideoTemp-o3. These differ from the separately supplied aggregate updates in Table 2; the diagnostic rows do not decompose those updates. Their corresponding per-question outputs are needed to reconcile the two sources.

Return diagnostics. Most questions return at the second observation, with higher accuracy than final fallback groups throughout Table 8. Post-stop changes show that early agreement need not persist with more frames. The return groups contain different questions selected by the policy, so

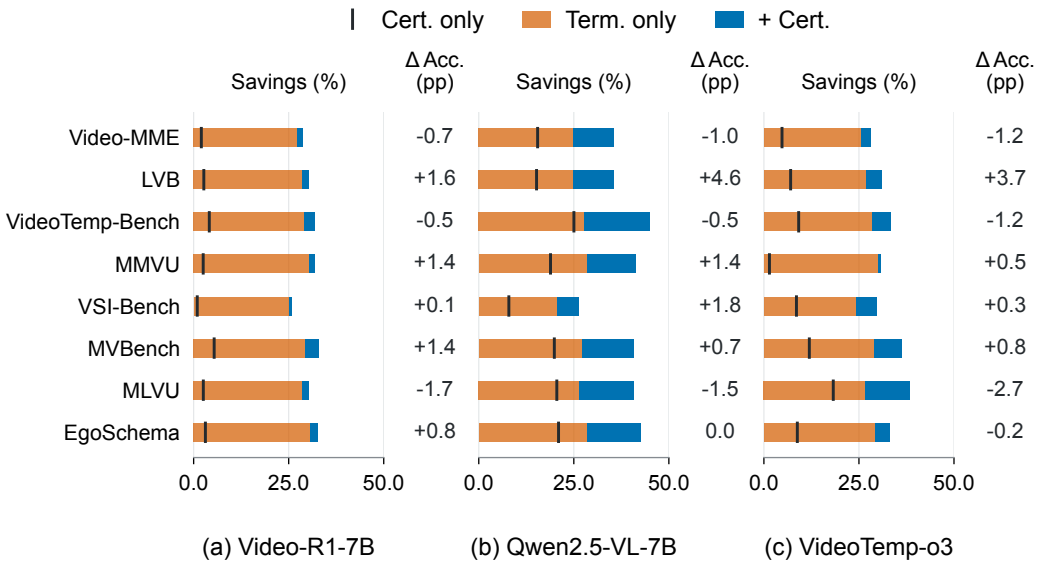


Figure 6: Trajectory savings relative to disabling both stopping rules: termination alone (orange), both rules (full bars), and certification alone (ticks). For MMVU and LVB, ΔAcc compares full RRV with the within-observation configuration in Table 3; other accuracy differences and all savings compare against no stopping in the archived experiment.

their accuracy gaps do not isolate the effect of stopping. Share uses all questions; Acc. uses each return group. Early certification counts stops before eight rounds over all $3n$ cached observations, including unvisited ones. Reversal measures disagreement with the full-budget 64-frame leader among $\mathcal{F}_2$ returns.

Frame sensitivity. Table 9 compares resampling at 16 frames with expansion to 32 frames using Video-R1, four trajectories per option order ($4 \times K$ total) in each condition, and a 2,048-token limit. Expansion changes answers more often on all three benchmarks, especially initially incorrect answers; all 95% intervals exclude zero. Correct answers also change, so expansion need not improve accuracy. With candidate counts fixed, expansion doubles the initial frames per trajectory; this comparison isolates changes in observation size from changes in the number of sampled answers, while increasing visual-input cost. Intervals use 10,000 video-cluster bootstrap replicates with seeds 2027, 2028, and 2029 for MVBench, MLVU, and LVB, respectively, recomputing subgroup denominators within the fixed trajectory cache, without multiplicity adjustment.

Table 9: Answer changes under resampling and observation expansion on Video-R1.

Benchmark	Initial-answer group	n	Answer-change rate (%)		Difference (pp) [95% CI]
			16 frames Resample	32 frames Expand	
MVBench	All	3,600	6.3	13.0	+6.7 [5.5, 7.8]
	Initially correct	2,312	4.1	9.0	+4.9 [3.8, 6.1]
	Initially incorrect	1,288	10.2	20.0	+9.8 [7.3, 12.3]
MLVU	All	2,174	3.4	14.2	+10.8 [9.3, 12.3]
	Initially correct	1,265	1.6	5.6	+4.0 [2.7, 5.4]
	Initially incorrect	909	5.9	26.2	+20.2 [17.4, 23.2]
LVB	All	1,337	3.7	14.3	+10.5 [8.7, 12.5]
	Initially correct	777	2.3	7.6	+5.3 [3.4, 7.3]
	Initially incorrect	560	5.7	23.6	+17.9 [14.4, 21.5]

Difference = expansion − resampling.

Table 8: Return diagnostics on the archived trajectory cache. MLVU rows do not decompose the separately supplied aggregate updates in Table 2.

Model	Benchmark	n	$\mathcal{F}_2$ agree		$\mathcal{F}_M$ agree		$\mathcal{F}_M$ fallback		Certificate	Reversal
			Share (%)	Acc. (%)	Share (%)	Acc. (%)	Share (%)	Acc. (%)	Early (%)	Rate (%)
Video-R1	Video-MME	900	81.7	52.1	10.4	45.7	7.9	36.6	16.2	10.1
	LVB	1337	85.6	62.4	9.1	38.5	5.3	35.2	21.7	7.4
	VTB	1200	87.5	78.5	6.7	68.8	5.8	41.4	32.2	4.9
	MMVU	625	90.9	69.2	5.3	33.3	3.8	29.2	20.1	4.8
	VSI-Bench	1124	76.2	33.5	13.9	33.3	9.9	28.8	6.9	14.5
	MVBench	3600	87.5	66.7	8.0	38.3	4.5	44.8	43.7	7.3
	MLVU	2174	85.5	64.3	8.7	60.0	5.8	42.9	20.4	8.7
	EgoSchema	500	91.8	61.2	5.4	29.6	2.8	14.3	24.7	4.6
Qwen2.5-VL	Video-MME	900	75.0	53.6	12.3	42.3	12.7	28.1	60.7	14.4
	LVB	1337	75.2	64.9	13.8	44.6	11.0	40.1	60.5	13.5
	VTB	1200	82.6	79.2	10.3	66.9	7.1	49.4	79.7	9.2
	MMVU	625	85.9	69.3	6.4	42.5	7.7	37.5	69.5	7.8
	VSI-Bench	1124	63.2	40.6	20.1	29.6	16.7	33.0	45.3	23.4
	MVBench	3600	81.2	69.3	9.6	43.5	9.2	37.8	70.0	10.6
	MLVU	2174	79.3	64.6	11.9	53.1	8.9	37.3	73.5	12.4
	EgoSchema	500	85.2	73.0	8.0	52.5	6.8	52.9	77.1	6.1
VideoTemp-o3	Video-MME	900	76.8	56.6	14.0	40.5	9.2	41.0	24.3	13.6
	LVB	1337	80.6	67.5	10.8	46.5	8.6	32.2	36.1	11.7
	VTB	1200	85.4	79.8	8.8	57.5	5.8	49.3	43.6	7.1
	MMVU	625	90.6	72.1	4.8	33.3	4.6	37.9	7.9	6.9
	VSI-Bench	1124	72.8	43.8	15.5	44.3	11.7	28.8	44.3	16.9
	MVBench	3600	86.3	70.9	7.8	39.5	5.9	35.8	53.8	8.4
	MLVU	2174	79.9	68.1	13.1	62.5	7.0	35.3	70.9	12.6
	EgoSchema	500	87.4	70.3	8.6	41.9	4.0	40.0	40.9	8.5

A.6 Returned-Answer Reliability

On the archived replay, c_{RRV} and s_+ rank the same returned answers. Averaged equally across the 24 model–benchmark pairs, their AUROCs are 71.6% and 71.5%, while their areas under the risk–coverage curve (AURC; lower is better) are 25.5% and 26.0%, respectively. The lower mean AURC indicates lower risk averaged over coverage levels for the fixed answer set; question-answering accuracy is unchanged by reranking. Improvements are not consistent: c_{RRV} reduces AUROC on Qwen2.5-VL and VideoTemp-o3. This diagnostic uses the original MLVU predictions described above; no confidence intervals were computed.

B QUALITATIVE ANALYSIS

Answer changes across observations. Figure 7 uses 40 unweighted votes over five option orders at each 16/32/64-frame observation, without early stopping. Qwen2.5-VL remains incorrect in the torsion test; in the reaction example, its answer changes from correct to incorrect and back.

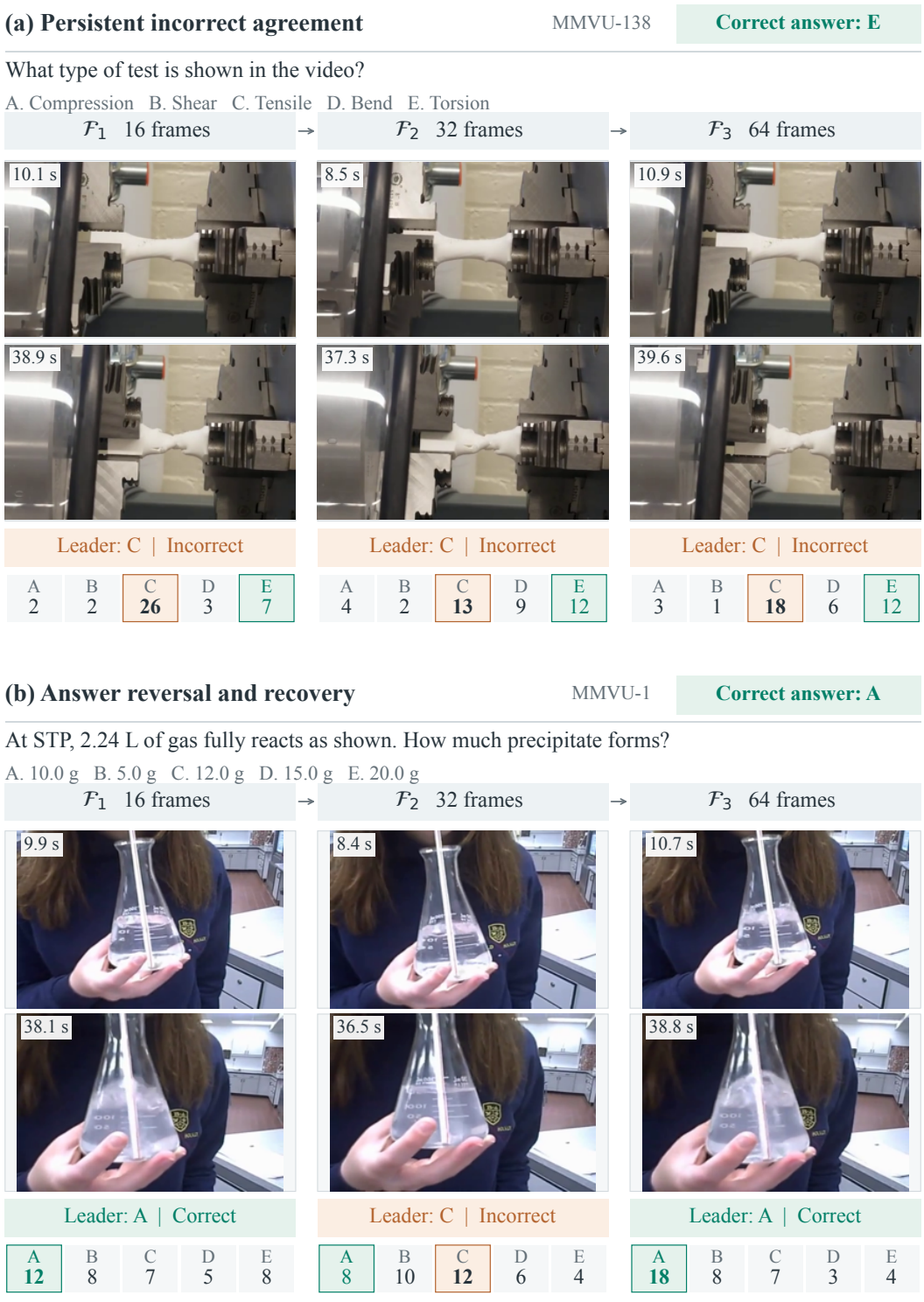


Figure 7: Qwen2.5-VL prediction paths and option vote counts on MMVU.

C Theoretical Details

The pathwise results require finite $\mathcal{O}_{\mathbf{q}}$ and an answer map fixed before generation. Answers and weights are fixed functions of trajectories and token probabilities, with $0 < \omega_i \leq 1$.

The population part of Proposition 2 additionally assumes trajectories i.i.d. across rounds within each presentation, independent presentations, and a unique leader $a_j^\star$ of P_j in Eq. (17) with margin $\Delta_j > 0$. Observations need not be independent.

C.1 Certification (Proposition 1)

Completion bounds and certificate. After r balanced rounds at $\mathcal{F}_j$, let $R = n_{\max} - r$, and for presentation ℓ let $A_{j,\ell}^{(r)}(o)$ and $W_{j,\ell}^{(r)}$ be the weight accumulated by answer o and in total. Completing the budget adds R trajectories of weight in $(0, 1]$ to every presentation, so the final support of o lies in $[L_{j,r}(o), U_{j,r}(o)]$ with:

$$L_{j,r}(o) = \frac{1}{K}\sum_{\ell=1}^{K} \frac{A_{j,\ell}^{(r)}(o)}{W_{j,\ell}^{(r)} + R}, \qquad U_{j,r}(o) = \frac{1}{K}\sum_{\ell=1}^{K} \frac{A_{j,\ell}^{(r)}(o) + R}{W_{j,\ell}^{(r)} + R}. \tag{14}$$

These are $\underline{p}_{j,r}$ and $\overline{p}_{j,r}$ in Eq. (9), and the certificate is:

$$B_{j,r} = \mathbf{1}\left[L_{j,r}(\hat{a}_j^{(r)}) > \max_{o \neq \hat{a}_j^{(r)}} U_{j,r}(o) \right], \tag{15}$$

where the maximum ranges over every canonical answer, including unseen ones with $A = 0$ (a dummy competitor with $A = 0$ if none exists), $B_{j,r} = 0$ when $\hat{a}_j^{(r)} = \perp$, ties follow the fixed order over $\mathcal{O}_{\mathbf{q}}$, and the implementation adds a 10^{-12} tolerance to the strict inequality. Sampling stops at the first round with $B_{j,r} = 1$ or at $r = n_{\max}$, and the leader at that round is $\hat{a}_j$.

Proof of Proposition 1. *Pathwise guarantee.* If $B_{j,r} = 1$, every completion gives $\hat{a}_j^{(r)}$ support at least $L_{j,r}(\hat{a}_j^{(r)})$ and every competitor o at most $U_{j,r}(o) < L_{j,r}(\hat{a}_j^{(r)})$, so $\hat{a}_j^{(r)}$ is the strict leader of every completion, including the full $n_{\max}$-round draw.

Firing condition. Let $c_{j,\ell}^{(r)}(o)$ count the first r trajectories of presentation ℓ answering o, so that $\bar{m}_{j,r} = \min_{o \neq \hat{a}} \frac{1}{Kr}\sum_\ell \big(c_{j,\ell}^{(r)}(\hat{a}) - c_{j,\ell}^{(r)}(o)\big)$ with $\hat{a} = \hat{a}_j^{(r)}$. Under unit weights, $A_{j,\ell}^{(r)}(o) = c_{j,\ell}^{(r)}(o)$ and $W_{j,\ell}^{(r)} = r$, so both bounds share the denominator $n_{\max}$ and Eq. (15) reads $\frac{1}{K}\sum_\ell c_{j,\ell}^{(r)}(\hat{a}) > \frac{1}{K}\sum_\ell c_{j,\ell}^{(r)}(o) + R$ for every $o \neq \hat{a}$, i.e. $r\,\bar{m}_{j,r} > n_{\max} - r$, which is Eq. (12); $\bar{m}_{j,r} \in [0, 1]$ since $0 \leq c_{j,\ell}^{(r)}(o) \leq r$. Under arbitrary weights, $A_{j,\ell}^{(r)}(\hat{a}) \leq W_{j,\ell}^{(r)} \leq r$ gives $L_{j,r}(\hat{a}) \leq r/n_{\max}$ while $U_{j,r}(o) \geq R/n_{\max}$, so $B_{j,r} = 1$ forces $r > R$, i.e. $r > n_{\max}/2$. □

Implementation scope. Two cases fall outside these assumptions: missing probabilities can reset all trajectories in a presentation to unit weights, revising earlier weights; post-hoc semantic grouping of open-ended answers does not fix the answer map before generation.

C.2 Cost of the Nested Schedule

Corollary 1 (cost). Let $c_j > 0$ be the per-trajectory cost at $\mathcal{F}_j$, r_j the rounds used there, $j^\star$ the return observation, and $\pi_j = \Pr(j^\star \geq j)$. Relative to $Kn_{\max}$ trajectories at $\mathcal{F}_M$, the cost ratio $\rho = \sum_{j \leq j^\star} r_j c_j / (n_{\max} c_M)$ satisfies:

$$\frac{(c_1 + c_2)(\lfloor n_{\max}/2 \rfloor + 1)}{n_{\max}\, c_M} \leq \rho \leq \frac{\sum_{j=1}^{M} c_j}{c_M}, \qquad \mathbb{E}[\rho] = \sum_{j=1}^{M} \frac{\pi_j\, \mathbb{E}[r_j \mid j^\star \geq j]\, c_j}{n_{\max}\, c_M}. \tag{16}$$

Proof. By Proposition 1, certification retains the full-budget leader and every visited observation uses $r_j \in [\lfloor n_{\max}/2 \rfloor + 1,\, n_{\max}]$ rounds. Each visited $j \leq j^\star$ costs Kr_jc_j, so dividing the total by $Kn_{\max}c_M$ gives ρ. The return rule visits at least $\mathcal{F}_1$ and $\mathcal{F}_2$, so the lower bound is attained when

$j^\star = 2$ and both certify at the earliest round, the upper bound when every observation exhausts its budget. The expectation follows from $\rho = \sum_j \mathbf{1}[j^\star \geq j]\, r_j c_j/(n_{\max} c_M)$ and $\mathbb{E}[\mathbf{1}[j^\star \geq j]\, r_j] = \pi_j\, \mathbb{E}[r_j \mid j^\star \geq j]$. For $c_j \equiv 1$ the lower bound $2(\lfloor n_{\max}/2 \rfloor + 1)/n_{\max}$ exceeds one for every $n_{\max} \geq 2$. For $c_j \propto b_j = (16, 32, 64)$ and $n_{\max} = 8$ the bounds are $15/32$ and $7/4$, and without certification $\mathbb{E}[\rho] = (b_1 + b_2)/b_M + \pi_3 = 3/4 + \pi_3$, which is below one iff $\pi_3 < 1/4$. □

C.3 Margin Stability (Lemma 1)

Proof of Lemma 1. Let $D = \max_o |p(o) - q(o)| < m_q/2$. For every $o \neq \hat{a}_q$, $p(\hat{a}_q) - p(o) \geq \big(q(\hat{a}_q) - D\big) - \big(q(o) + D\big) \geq m_q - 2D > 0$, so $\hat{a}_q$ is the unique leader of p and the tie rule is not invoked. The statement needs $|\mathcal{O}_\mathbf{q}| \geq 2$ and a valid leader of q; it holds for any two supports, whether they differ by weighting, by sampling, or by observation. □

C.4 Early Return (Proposition 2)

Population support. Renormalized top-20 entropy is at most $\ln 20$, so entropy weights satisfy $\omega_i = \exp(-U_i/\sigma) \geq \omega_{\text{low}} = \exp(-\ln 20/\sigma) > 0$. The same lower bound holds for unit weights. Let (ω, a) be the weight and answer of a trajectory from $x_{j,\ell}$. The population support at $\mathcal{F}_j$ is:

$$P_j(o) = \frac{1}{K} \sum_{\ell=1}^{K} \frac{\mathbb{E}[\omega\, \mathbf{1}(a = o) \mid x_{j,\ell}]}{\mathbb{E}[\omega \mid x_{j,\ell}]}, \tag{17}$$

the almost-sure limit of $p_j^{(r)}$ as $r \to \infty$ by the strong law of large numbers, since every denominator is at least $\omega_{\text{low}} > 0$.

Proof of Proposition 2. *Pathwise part.* Lemma 1 with $q = p_j$ and $p = p_M$ gives the margin condition. If $p_M(\hat{a}_j) > 1/2$, every competitor has support at most $1 - p_M(\hat{a}_j) < 1/2$ because supports sum to at most one, so $\hat{a}_j$ still leads; an overturn therefore requires $p_M(\hat{a}_j) \leq 1/2$, a fall of at least $p_j(\hat{a}_j) - 1/2 \geq c_{\text{RRV}} - 1/2$, since $p_j(\hat{a}_j) \geq c_{\text{RRV}}$ by Eq. (10).

Population part. If $\hat{a}_j = a_j^\star$ and $a_j^\star = a_M^\star$, then $\hat{a}_j = a_M^\star$; hence $\{\hat{a}_j \neq a_M^\star\} \subseteq \{a_j^\star \neq a_M^\star\} \cup \{\hat{a}_j \neq a_j^\star\}$. Lemma 1 with $q = P_j$ and $p = P_M$ shows that the first event requires $\max_o |P_M(o) - P_j(o)| \geq \Delta_j/2$. For the second, whether sampling at $\mathcal{F}_j$ stops early or runs to $n_{\max}$, $\hat{a}_j$ is the leader of the completed $n_{\max}$-round draw by Proposition 1, so it suffices to bound that draw. Under unit weights, for a competitor $o \neq a_j^\star$ the difference $p_j(o) - p_j(a_j^\star)$ is the mean of $K n_{\max}$ independent variables in $[-1, 1]$ with mean $P_j(o) - P_j(a_j^\star) \leq -\Delta_j$, so Hoeffding's inequality gives $\Pr\big(p_j(o) \geq p_j(a_j^\star)\big) \leq e^{-K n_{\max} \Delta_j^2/2}$, counting ties as errors; a union bound over the $|\mathcal{O}_\mathbf{q}| - 1$ competitors bounds $\Pr(\hat{a}_j \neq a_j^\star)$, and the union of the two events gives Eq. (13). The first term is a function of P_j and P_M alone, hence independent of K, $n_{\max}$, and σ.

Spurious agreement. If $a_{j-1}^\star \neq a_j^\star$, then $\hat{a}_{j-1} = \hat{a}_j$ requires $\hat{a}_{j-1} \neq a_{j-1}^\star$ or $\hat{a}_j \neq a_j^\star$, so $\Pr(\hat{a}_{j-1} = \hat{a}_j)$ is at most the sum of the two sampling terms. □