COG-FOVEA: CONDITIONAL VISUAL ACQUISITION AND ANSWER SELECTION FOR VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse frame sampling can omit brief events essential for video reasoning. Extending textual reasoning may improve the interpretation of observed frames, but does not acquire missing observations. We introduce Cog-Fovea, an inference framework with frozen weights that separates two decisions: when to inspect additional video frames and when to replace an existing answer. An initial global pass at high resolution produces an answer and a confidence score. Queries with low confidence trigger temporal grounding and a hybrid view at lower resolution combining local samples with global anchors. Arbitration based on confidence selects between the original and hybrid answers, retaining the initial answer as a candidate rather than automatically accepting the new prediction. All stages share a frozen backbone, with a cap on visual tokens per pass and total computation accounting for routed calls. A conditional analysis of event coverage characterizes when localized sampling can capture brief events. Experiments on five benchmarks examine accuracy, inference cost, and answer selection behavior. On Video-MMMU, Cog-Fovea achieves 60.2% accuracy, compared with 57.0% using only the global pass. Inference without explanations averages 6.3 seconds per query. Approximately 35% of queries receive additional inspection. Arbitration provides a numerical gain of 0.6 percentage points over accepting every available hybrid answer, but reduces accuracy on a diagnostic subset with narrow temporal windows. These findings support conditional visual acquisition as a strategy for allocating computation at test time and show why acquiring more evidence and accepting a revised answer must be evaluated separately.

1 INTRODUCTION

Test-time scaling improves the capabilities of large language models by allocating additional computation to reasoning, for example through chain-of-thought generation (Jaech et al., 2024; Guo et al., 2025). For video, however, the inference budget determines not only how long a model reasons but also which observations are available to that reasoning process. Video-capable models (Li et al., 2025c; Bai et al., 2025) commonly operate on a restricted sample of frames. If a brief event falls between those frames, extending the textual computation cannot directly recover the missing observation. Further reasoning may still improve an answer by exploiting partial evidence and prior knowledge, but it does not provide another view of the source video.

This distinction motivates treating observation acquisition as an explicit inference decision. Figure 1 illustrates a brief slip near a warning cone that can be absent from a sparse sample showing only walking before and after the event. Reasoning over that sample and inspecting the omitted interval are different operations. Under the conditional Markov assumption in Section 3.1, the data processing inequality formalizes the information restriction imposed by a fixed visual context. It does not imply that longer reasoning cannot improve accuracy, or that additional frames necessarily yield a better answer. The latter qualification matters under a limited visual context: increasing temporal coverage may require lower spatial resolution and make small details harder to recognize.

Recent work already makes perception adaptive. *Select Less, Reason More* (Li et al., 2026b) learns evidence-aware frame selection with localized resampling, while APPO (Du et al., 2026) uses attention-guided policy optimization to improve perception. Training is not the only route to adaptive sampling: *Divide, then Ground* (DIG) (Li et al., 2026a) switches between uniform and

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.

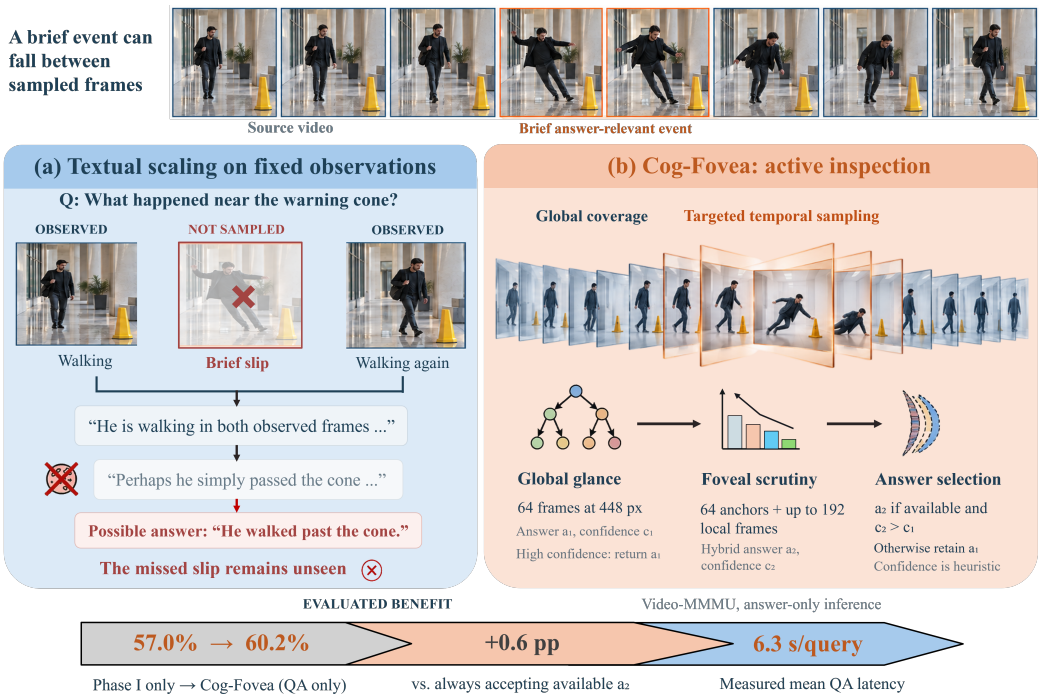


Figure 1: Observation acquisition and answer selection in Cog-Fovea. Top: the illustrative slip near a warning cone can fall between sampled frames. (a) Additional textual reasoning operates on the available observations. (b) Low initial-answer confidence triggers inspection of a predicted interval, together with global anchors, and a confidence comparison determines whether to replace the initial answer. The scene is a sampling illustration rather than an experimental prediction trace. Bottom: on the full Video-MMMU set, Phase I-only answering records 57.0% accuracy and Cog-Fovea records 60.2%. The separate arbitration comparison gives a numerical gain of 0.6 percentage points over accepting every available hybrid answer. Mean QA-only latency is 6.3 s per query (Tables 6 and 2).

query-aware selection according to whether a question is global or localized. These approaches motivate a complementary question: after a frozen model has produced an initial answer, when should it acquire a different visual view, and when should that view change the answer? The decision to inspect and the decision to replace need not coincide, because a temporally denser but spatially coarser view can resolve one ambiguity while introducing another.

Cog-Fovea (Cognitive Foveated Video Reasoning) addresses these decisions with a three-phase inference policy. Phase I generates a concise answer from 64 global frames at nominal 448 px resolution. Low initial-answer confidence activates Phase II, which predicts a temporal interval and, when local source frames are available, constructs a hybrid view from up to 192 local selections and 64 global anchors at nominal 224 px. Overlapping source frames are deduplicated. The visual-token limit applies separately to each pass, not to the sum of computation over the query. Phase III retains the initial answer unless an available hybrid answer has a higher confidence score. This comparison is a heuristic for answer selection, not a verification that missing evidence has been recovered. Both views use the same frozen VideoAuto-R1 backbone (Liu et al., 2026), rather than separately trained experts. Optional explanations are generated after the answer has been selected and cannot alter it.

The contribution is the coupling of conditional cross-resolution acquisition with selective answer replacement, rather than active perception in isolation. The initial high-resolution answer remains a candidate even after the model inspects a denser temporal view. This design makes the trade-off between obtaining new observations and retaining an existing answer explicit, without training another selector or updating the backbone. The event-coverage analysis characterizes the probability of sampling a brief event under stated assumptions about timing and window coverage, rather than providing an unconditional accuracy guarantee.

2 Related Work

2.1 Reasoning and Conditional Test-Time Computation

Video-R1 (Feng et al., 2026), Video-RTS (Wang et al., 2025), and VideoChat-R1.5 (Yan et al., 2026) extend reinforcement learning and test-time reasoning to video understanding. VideoAuto-R1 (Liu et al., 2026) uses initial-answer confidence to determine whether to continue textual reasoning. Cog-Fovea retains confidence as a routing signal but changes the routed action to acquiring another visual context, followed by selection between answers from different spatial–temporal views. This separates two dimensions of inference-time computation: reasoning over observations and changing the observations themselves. The dimensions can be interleaved, rather than requiring perception to finish before reasoning begins. Additional reasoning can resolve ambiguity or clarify temporal relations within a fixed context, while further sampling can expose evidence outside that context.

Perceptual reliability also depends on how the model uses the observations it already has. SEASON (Wu et al., 2026) addresses temporal hallucination through token-level self-diagnosis and adaptive contrastive decoding. In *Deeper Thought, Weaker Aim*, Peng et al. (2026) study visual-attention dispersion during extended multimodal reasoning and introduce training-free Visual Region-Guided Attention to reweight selected attention heads. These interventions operate on the interpretation of available visual input. Cog-Fovea instead controls access to additional source frames and whether to accept the resulting answer. Attention drift over observed content and omission of an event during sampling are distinct failure modes, so neither perspective subsumes the other.

2.2 Evidence Selection and Spatial–Temporal Allocation

Learning which frames matter is a common approach to efficient video reasoning. *Select Less, Reason More* (Li et al., 2026b) couples evidence-aware reinforcement learning with frame selection and localized resampling. FrameOracle (Li et al., 2025a) predicts both frame relevance and how many frames to retain, using a curriculum from proxy signals to keyframe supervision. Qin et al. (2026) train a lightweight selector in two stages: frame–query relevance pretraining followed by reinforcement learning with rewards at both frame and frame-combination levels. The latter stage targets frames that improve downstream task performance, rather than relevance alone. Cog-Fovea does not learn a new selection module. It uses the frozen backbone's grounding capability within a fixed confidence-routing and answer-selection policy.

Adaptive selection can also be training-free. DIG (Li et al., 2026a) distinguishes global from localized questions, applying uniform sampling to the former and a query-aware selection pipeline to the latter. Its conditional computation is organized around query type. Cog-Fovea instead routes after producing an initial answer, using that answer's confidence, and retains the high-resolution answer for comparison with the lower-resolution hybrid candidate. The distinction is thus the routing signal and the coupling of observation acquisition with answer retention, not the absence of training alone.

Other methods improve visual evidence use during training. APPO (Du et al., 2026) derives fine-grained policy-optimization signals from attention to critical frames, while VisionCoach (Lee et al., 2026) selectively introduces visual prompts during training and uses self-distillation to support inference without those prompts. These approaches complement inference-time control with fixed weights. For Cog-Fovea, spatial readability remains relevant even when temporal selection succeeds. Denser sampling does not guarantee that small objects or text survive downsampling, which motivates keeping the answer from the high-resolution view available.

2.3 Temporal Grounding and Active Visual Inspection

Thinking with Videos (Zhang et al., 2026) develops tool-augmented video reasoning, and Time-R1 (Wang et al., 2026) learns temporal grounding. VideoAgent (Wang et al., 2024) organizes iterative retrieval of evidence from long videos. Training-free temporal grounding (Zheng et al., 2024) further shows that localization can be implemented without additional task-specific weight updates. Related image-based approaches, including DeepEyes (Zheng et al., 2026) and ZoomEye (Shen et al., 2025), acquire finer visual evidence through active inspection. LOVE-R1 (Fu et al., 2025b) applies adaptive zoom-in reasoning to long-video understanding.

The available visual actions also differ across reasoning frameworks. Act2See (Ma et al., 2026b) learns to interleave textual reasoning with retrieval of source frames or synthesis of hypothetical frames. ReRe (Ma et al., 2026a) provides a training-free approach to spatial reasoning that checks an initial hypothesis using novel-view videos rendered from predicted geometry. Both integrate visual inspection into reasoning, but their action spaces include generated visual content. Cog-Fovea restricts acquisition to frames from the source video, uses one temporal sampling interval rather than iterative multi-tool search, and enforces a visual-token cap per pass.

Across these lines of work, adaptive perception is an established strategy rather than a novelty claim on its own. Cog-Fovea combines high-resolution global answering, conditional lower-resolution temporal resampling, and confidence-based answer retention in one frozen model. The hybrid input reuses global anchors at lower resolution, while the answer obtained from the original high-resolution input remains available. Acquisition changes the evidence presented to the backbone, and arbitration chooses among the answers using their context-specific confidence scores. This separation exposes the spatial–temporal trade-off without assuming that the denser view or the higher-scoring candidate is always correct.

3 METHODOLOGY

Cog-Fovea treats test-time video reasoning as two coupled decisions: whether to acquire further observations and whether to replace the current answer. A single frozen backbone supports both decisions through different visual inputs, without training separate spatial and temporal experts. Figure 2 summarizes the resulting policy. Phase I answers from 64 globally sampled frames at nominal 448 px resolution. Phase II is invoked only when the initial confidence falls below a specified threshold, or equals it. It first predicts a temporal interval and then, if local source frames exist, answers from a hybrid view combining global anchors and local samples at nominal 224 px resolution. Phase III selects between the available answers without another model call. Thus, the three conceptual phases are not three unconditional backbone evaluations: temporal grounding and hybrid answering are separate calls within Phase II. Panel A shows this control flow, Panel B the hybrid input, Panel C the decision rules, and Panel D the spatial–temporal allocation and compute accounting. Each visual prefill obeys a cap of $B = 16{,}384$ tokens, rather than sharing a single cap across the entire query. The equations below assume successful model calls and nonempty answer outputs. Optional explanations are generated only after selection and lie outside the answer-only pipeline.

3.1 OBSERVATION ACQUISITION AND TEMPORAL COVERAGE

Fixed-context reasoning versus new observations. Let $\mathcal{X}$ denote a video of duration $\mathcal{T}_{total} > 0$, q its question, A the correct answer, and θ the frozen model parameters. For the analysis, videos, questions, and answers are random variables under the evaluation distribution, with mutual information denoted by $\mathcal{I}$. Once the visual context $\mathcal{V}_{global}$ and prompting policy are fixed, generating a reasoning trace z and answer a does not access the source video again. If decoding receives no additional answer-dependent input or randomness, then $A \to \mathcal{V}_{global} \to (z, a)$ forms a Markov chain conditional on (q, θ). The data processing inequality consequently implies

$$\mathcal{I}(A; z, a \mid q, \theta) \le \mathcal{I}(A; \mathcal{V}_{global} \mid q, \theta). \tag{1}$$

This statement limits information in the output, not the accuracy improvements possible through better reasoning over existing observations. In contrast, adding observations $\mathcal{V}'$ to the retained global context gives the chain-rule identity

$$\begin{aligned} &\mathcal{I}(A; \mathcal{V}_{global}, \mathcal{V}' \mid q, \theta) \\ &\quad - \mathcal{I}(A; \mathcal{V}_{global} \mid q, \theta) \\ &= \mathcal{I}(A; \mathcal{V}' \mid q, \theta, \mathcal{V}_{global}) \ge 0. \end{aligned} \tag{2}$$

The additional observations may be chosen adaptively from the initial context, and their conditional information can still be zero. This identity applies to augmentation of the original observations, not replacement by a compressed view. Cog-Fovea uses a predicted temporal window to choose observations, but neither estimates nor optimizes mutual information. Because its hybrid pass re-encodes even the global anchors at lower resolution, that pass is not guaranteed to contain more answer-relevant information than the original high-resolution input. Appendix C.1 analyzes when

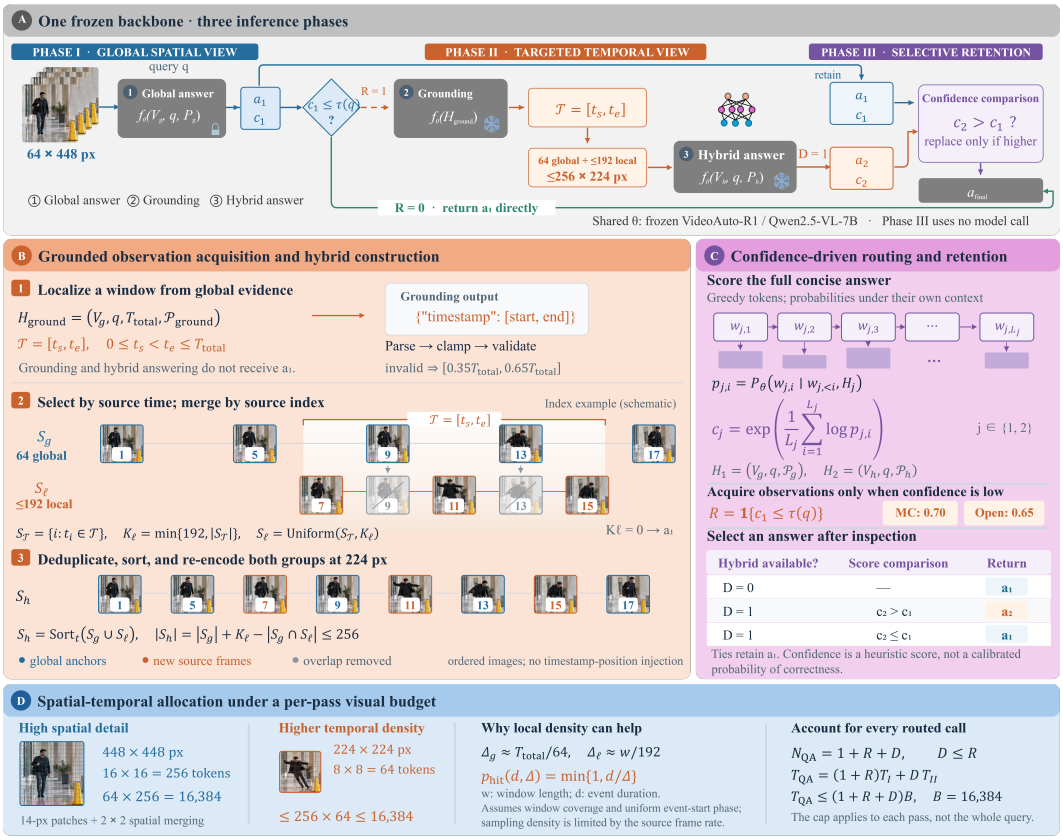


Figure 2: Cog-Fovea architecture and answer-only inference. (A) A shared frozen backbone generates the global answer, performs conditional temporal grounding, and generates the hybrid answer when local frames are available. These model calls are distinct from the three inference phases, with Phase III requiring no model call. (B) Global anchors and local selections are deduplicated by source index, ordered chronologically, and re-encoded at 224 px. Frame indices are schematic. (C) Token probabilities define heuristic confidence scores for routing and answer selection. The initial answer is retained when no hybrid answer is available or its confidence does not exceed the initial score. (D) Spatial–temporal allocation, conditional event coverage, and total visual-token accounting under the per-pass cap $B = 16{,}384$. Optional post-selection explanation calls are not shown. In the diagram, V_g, V_h, P_g, and P_h abbreviate $\mathcal{V}_{global}$, $\mathcal{V}_{hybrid}$, $\mathcal{P}_{global}$, and $\mathcal{P}_{foveal_hypo}$, respectively.

localized sampling can capture brief events under explicit timing and coverage assumptions. This analysis concerns observation availability rather than guaranteed improvements in answer accuracy.

3.2 Phase I: Confidence-Based Routing

The global-answer call applies $\mathcal{P}_{global}$ to $K_{anchor} = 64$ uniformly sampled frames and returns a concise hypothesis a_1. Table 3 in Appendix A gives the prompts. For each nonempty candidate a_j, $j \in \{1, 2\}$, let $(w_{j,1}, \ldots, w_{j,L_j})$ be its hypothesis tokens. We compute

$$c_j = \exp\left(\frac{1}{L_j}\sum_{i=1}^{L_j} \log P_\theta(w_{j,i} \mid w_{j,<i}, H_j)\right), \tag{3}$$

with contexts $H_1 = (\mathcal{V}_{global}, q, \mathcal{P}_{global})$ and $H_2 = (\mathcal{V}_{hybrid}, q, \mathcal{P}_{foveal_hypo})$. Here, $w_{j,<i}$ denotes the generated prefix and L_j counts the scored tokens. The score uses the full hypothesis and excludes any subsequent rationale or reflection. It is the geometric mean of token probabilities under the model, not a calibrated probability of answer correctness. Decoding is greedy, which fixes token selection without making the selected tokens' model probabilities equal to one.

For an answer-format-specific threshold $\tau(q)$, the routing decision and observed routing fraction over N evaluation queries are

$$R = \mathbf{1}\{c_1 \le \tau(q)\}, \qquad \gamma = \frac{1}{N}\sum_{n=1}^{N} R_n. \tag{4}$$

When $R = 0$, inference stops with a_1. When $R = 1$, including equality at the threshold, the policy requests additional inspection. The thresholds are $\tau_{mc} = 0.70$ for multiple-choice questions and $\tau_{open} = 0.65$ for open-ended questions. The approximately 35% routing fraction on Video-MMMU results from these thresholds, rather than from enforcing a routing quota. Normalizing by hypothesis length reduces the score's direct dependence on length, but does not remove its sensitivity to wording or predictable continuations. Confidence therefore serves as a routing heuristic, not a test of whether the sampled input contains sufficient evidence.

3.3 Phase II: Temporal Localization and Hybrid Sampling

Proposing an inspection interval. A routed query first invokes temporal grounding on the global frames, question, and video duration:

$$\begin{aligned} H_{ground} &= (\mathcal{V}_{global}, q, \mathcal{T}_{total}, \mathcal{P}_{ground}), \\ \widehat{\mathcal{T}} &= \mathrm{Parse}(\mathrm{Decode}(P_\theta(\cdot \mid H_{ground}))) \,. \end{aligned} \tag{5}$$

The parser extracts a start–end pair $\widehat{\mathcal{T}}$ from the generated JavaScript Object Notation (JSON) output, returning $\perp$ when extraction fails. The proposal becomes a sampling interval $\mathcal{T} = [t_{start}, t_{end}]$ after validation or fallback. Grounding does not receive the initial answer a_1, so its output proposes where to inspect rather than assessing whether a_1 is correct. Cues before or after an event can suggest its location without resolving the question, although the pretrained model's grounding capabilities (Bai et al., 2025) do not guarantee sufficient evidence.

Finite predicted boundaries are clipped to $[0, \mathcal{T}_{total}]$ and accepted only if $t_{start} < t_{end}$ after clipping. An unparseable output or invalid pair instead selects the central 30% of the video, $[0.35\mathcal{T}_{total}, 0.65\mathcal{T}_{total}]$. This fallback specifies a sampling location, not a guarantee of relevance.

Constructing the hybrid view. Write x_i for source frame i, t_i for its timestamp, and $\mathcal{S}_g$ for the distinct source indices of the global anchors. As depicted in Figure 2B, local candidates, selected local indices, and the ordered hybrid sequence are

$$\begin{aligned} \mathcal{S}_{\mathcal{T}} &= \{i : t_i \in [t_{start}, t_{end}]\}, \\ K_\ell &= \min\{192, |\mathcal{S}_{\mathcal{T}}|\}, \\ \mathcal{S}_\ell &= \mathrm{Uniform}(\mathcal{S}_{\mathcal{T}}, K_\ell), \\ \mathcal{S}_h &= \mathrm{Sort}_t(\mathcal{S}_g \cup \mathcal{S}_\ell). \end{aligned} \tag{6}$$

The operator $\mathrm{Uniform}$ subsamples K_ℓ distinct source indices. It does not choose a random window, retains every candidate when fewer than 192 are available, and returns the empty set when $K_\ell = 0$. The operator Sort_t orders the union into $\mathcal{S}_h = (i_1, \ldots, i_m)$ by source timestamp, with $m = |\mathcal{S}_h|$. Timestamps determine membership in the interval, whereas source indices determine whether two selections refer to the same frame. The procedure neither expands a short window nor pads it with duplicate frames. If no local frames are available, the pipeline returns a_1 without a_2.

Otherwise, both anchor and local frames in $(x_{i_1}, \ldots, x_{i_m})$ are processed at nominal 224 px resolution, compared with 448 px in Phase I. Taking the union before sorting removes overlap with the anchors, giving $|\mathcal{S}_h| = |\mathcal{S}_g| + K_\ell - |\mathcal{S}_g \cap \mathcal{S}_\ell| \le 256$. Only $|\mathcal{S}_\ell \setminus \mathcal{S}_g|$ selections are new source frames, and this number can be less than K_ℓ. If every local selection is already an anchor, the hybrid call still runs on the lower-resolution anchor view but acquires no new source frames. The anchors preserve sparse observations outside the proposed interval, while reducing resolution accommodates more frames under the per-pass token cap. A broad interval approaches whole-video dense sampling, whereas a short interval may contribute few distinct observations. Deduplication can therefore produce fewer than 256 input images without padding.

The hybrid-answer call uses $\mathcal{P}_{foveal_hypo}$ to generate a concise a_2, and Equation 3 evaluates its score c_2 under H_2. It does not receive the text of a_1: the new hypothesis is generated separately, although

its visual input depends on the preceding routing and grounding decisions. Each candidate is scored under its own prompt and visual view, without rescoring both candidates in a shared context. Both groups of frames are re-encoded at lower resolution to respect the combined per-pass budget. They are supplied as individual images in chronological order, without explicit timestamp-position injection. Because the hybrid sequence is temporally nonuniform, adjacency preserves order but does not specify elapsed time. Using timestamps for retrieval and sorting is thus distinct from supplying timestamps as positional inputs.

3.4 Phase III: Confidence-Based Answer Selection

Let D indicate whether the routed hybrid-answer call completes. Under the assumption, $D = 1$ for routed queries with $K_\ell > 0$ and $D = 0$ for early exits or empty local candidate sets, so $D \le R$. The acceptance indicator S and returned answer implement the decision rule in Figure 2C:

$$S = \begin{cases} \mathbf{1}\{c_2 > c_1\}, & D = 1, \\ 0, & D = 0, \end{cases} \qquad a_{final} = \begin{cases} a_2, & S = 1, \\ a_1, & S = 0. \end{cases} \tag{7}$$

This rule is defined even when no c_2 exists, and a score tie keeps a_1. On completed routes, it compares uncalibrated scores from different prompts and visual contexts. It is neither a Bayesian update nor a classifier distinguishing spatial from temporal failures. A poorly localized view may still elicit high confidence, so retaining an alternative answer does not guarantee correctness.

Separating inspection gains from arbitration gains. For benchmarks that average per-query scores $u(a, A)$, the change from always returning the initial answer is exactly

$$\Delta U = \frac{1}{N} \sum_{n:S_n=1} \big[u(a_{2,n}, A_n) - u(a_{1,n}, A_n)\big]. \tag{8}$$

With exact-match accuracy, u is the correctness indicator. Only accepted replacements affect this difference, and each replacement may either fix an error or introduce one. Neither the reference answer A nor its score u is accessible to the inference policy. These identities compare policies using the same generated candidates and evaluation queries. For $u \in [0, 1]$, multiplying a score difference by 100 converts it to percentage points.

A separate comparison isolates arbitration by taking the reference policy that accepts every available a_2 and returns a_1 only when $D = 0$. The difference from that reference is

$$\Delta U_{\text{arb}} = \frac{1}{N} \sum_{\substack{n:D_n=1 \\ S_n=0}} \big[u(a_{1,n}, A_n) - u(a_{2,n}, A_n)\big]. \tag{9}$$

Equation 8 measures the full routed policy against initial answers, whereas Equation 9 attributes changes specifically to rejected second answers. The latter can be negative on a subset when reversions discard more correct answers than they restore.

Localization, hypothesis generation, and selection have distinct roles. The grounding output specifies an interval rather than an extended answer justification, and the following call answers from the acquired hybrid input. Selection then operates outside autoregressive generation using the two scores. Retaining a_1 makes it possible to preserve information such as small text or object details that resizing may remove from the hybrid view. However, the confidence comparison does not explicitly detect these losses.

Optional explanations are generated after answer selection and cannot revise the selected answer. Appendix C.2 details these calls and the computation accounting, including repeated visual prefills under the per-pass token cap.

4 Experiments

The experiments examine whether conditional observation acquisition and selective answer replacement offer a favorable accuracy–cost trade-off compared with direct answering and unconditional

Table 1: Benchmark scores (%). External rows follow the cited papers' protocols, and VideoAuto-R1 denotes the published auto-think reference. These descriptive comparisons are distinct from the always-think timing reference in Table 2.

Model	Mechanism	Reasoning & Fine-grained		Perception & Long-Video	
		MVP (Pairwise)	Video-MMMU	MVBench	VideoMME
Qwen2.5-VL (7B) Bai et al. (2025)	Direct Answer	36.5	54.7	67.1	66.0
Video-R1 (7B) Feng et al. (2026)	Always-Think	33.0	51.4	65.5	61.8
VideoChat-R1.5 (7B) Yan et al. (2026)	Iterative Perception	38.6	49.6	70.6	65.2
Thinking with Videos (7B) Zhang et al. (2026)	Tool-Augmented Reasoning	-	54.2	-	64.1
VideoAuto-R1 (7B) Liu et al. (2026)[a]	Reported Auto-Think	39.4	58.6	71.0	67.3
Cog-Fovea (7B)	**Active Scaling**	**40.7**	**60.2**	**71.8**	**68.6**

[a] VideoAuto-R1's published auto-think scores are 58.6% on Video-MMMU and 39.4% on MVP. The corresponding always-think scores are 58.7% and 39.8% Liu et al. (2026). The separate 8.2 s timing reference is always-think, not auto-think.

additional computation. We distinguish benchmark-level comparisons with published systems from controls targeting individual inference decisions. Direct concise answering removes local inspection, alternative sampling changes the visual input, and the always-accept policy tests arbitration on the same routed candidates. Together, these comparisons probe prompting, input allocation, and answer selection, while retaining the distinction between controls that isolate a decision and those that change several factors.

Experimental setup. All Cog-Fovea evaluations use a frozen VideoAuto-R1 checkpoint built on Qwen2.5-VL-7B (Liu et al., 2026; Bai et al., 2025). The benchmarks are MVP (Krojer et al., 2025), Video-MMMU (Hu et al., 2026), MVBench (Li et al., 2024), VideoMME (Fu et al., 2025a), and NExT-GQA (Xiao et al., 2024). Inference runs on A100 80 GB hardware with batch size one and seed 42. On NExT-GQA, internal QA comparisons use WUPS@0.9, while grounding is evaluated separately with mIoU and early exit disabled. Published results retain their original protocols and metrics. Appendix D provides the complete experimental setup, metric definitions, token statistics, and computation accounting.

4.1 Accuracy and Inference Cost

Benchmark-level comparisons. Cog-Fovea reaches 60.2% on Video-MMMU, compared with the published VideoAuto-R1 auto-think score of 58.6% (Table 1). The difference is 1.6 percentage points, with corresponding differences of 1.3, 0.8, and 1.3 points on MVP, MVBench, and VideoMME. These are descriptive differences between point estimates, not established statistically significant gains. Because the external systems follow their own evaluation protocols, these comparisons do not isolate the controller's contribution under a common protocol.

Dense answering and prompt choice. The 256-frame uniform dense concise-answer control records 57.8% accuracy at approximately 6.4 s per query, compared with 60.2% at 6.3 s for Cog-Fovea. This is a 2.4-point accuracy difference under the evaluated configurations. The 0.1 s latency difference is too small, on the available point estimates alone, to establish a reliable speed advantage. With the dense input held fixed, replacing the global prompt by the foveal prompt raises Video-MMMU accuracy to 58.9% (Table 5), leaving a 1.3-point difference from Cog-Fovea. The two dense controls compare prompt choices. Their comparisons with Cog-Fovea additionally change routing, visual allocation, and answer selection, so neither isolates the effect of localized sampling.

Answer-only versus explanation-enabled inference. QA-only Cog-Fovea averages 6.3 s per query against 8.2 s for the always-think timing reference, a reduction of approximately 23% (Table 2). The always-think row pairs this local timing record with the separately published accuracy of 58.7%. Its numerical accuracy difference from Cog-Fovea is therefore 1.5 points, distinct from the 1.6-point comparison with auto-think.

Early-exit and routed Video-MMMU queries have reported final-answer accuracies of 70.0% and 42.0%, respectively. The routed figure includes arbitration and is not the accuracy of the hybrid hypothesis alone. Because confidence selects these populations, their accuracies cannot by themselves measure the benefit of further inspection. Combining the reported branch scores and routing fraction

Table 2: Full-set Video-MMMU accuracy and average per-query inference cost. Cog-Fovea uses $\gamma \approx 0.35$. Relative latency is normalized by the 8.2 s always-think reference, whose accuracy is the published score. Cog-Fovea latency entries are the reported overall averages. Full and answer-only modes use the same answer-selection rule. Lower relative latency indicates faster inference.

Method	Passes	Text Tokens/query	Relative Latency	Latency (s/query)	Video-MMMU (%)
VideoAuto-R1 (always-think timing record)	1	∼149	1.0×	8.2	58.7
Uniform dense concise answer (256 frames)	1	∼45	0.78×	≈ 6.4	57.8
Self-Consistency (Naive, 5 samples, indep. prefills)	5	∼745	5.0×	41.0	58.9
Cog-Fovea (avg, γ=0.35, full)	∼2.1	∼155	1.1×	9.0	60.2
Cog-Fovea (avg, γ=0.35, QA-only)	**∼1.7**	**∼70**	**0.77×**	**6.3**	**60.2**

The Passes column counts model calls, not the three conceptual phases. Tokens/query counts generated tokens across the calls included in each configuration. Cog-Fovea values are averaged over early-exit and routed queries. Optional explanation calls are included but cannot change the selected answer.

gives $0.65 \times 70.0 + 0.35 \times 42.0 = 60.2\%$. This consistency check uses rounded summaries and does not reconstruct integer correctness counts.

Full mode retains 60.2% accuracy but averages 9.0 s per query. The unchanged answer score follows from fixing the answer before generating explanations, rather than demonstrating an accuracy benefit from reflection. At 9.0 s, full mode is slower than the 8.2 s always-think reference, so the approximately 23% reduction applies only to QA-only inference. The naive five-sample self-consistency control obtains 58.9% at 41.0 s, versus 60.2% at 6.3 s for QA-only Cog-Fovea. That control uses independent prefills, and this comparison should not be extended to optimized implementations that share caches. Appendix B provides the grounding comparisons, alternative-sampling controls, and component ablations.

4.2 When Confidence Arbitration Rejects Correct Answers

The narrow-window diagnostic evaluates the same 69 routed Video-MMMU queries under three answer-selection policies (Table 6, Panel C, in Appendix B). Each query has a nonempty local candidate set with fewer than 192 source frames, satisfying $0 < |\mathcal{S}_\mathcal{T}| < 192$ in Equation 6. All percentages use these 69 queries as the denominator, not the complete routed population. The initial answer, hybrid answer, and selected answer are correct on 33, 37, and 35 queries, giving accuracies of 47.8%, 53.6%, and 50.7%. Hybrid answering adds four correct predictions in aggregate relative to Phase I, but arbitration retains a net gain of only two. These differences count net changes in correctness rather than individual error corrections.

Of 43 rejected hybrid candidates, 13 are correct answers replaced by incorrect initial answers, while 11 are incorrect answers replaced by correct initial answers. The other 19 reversions leave correctness unchanged. Consequently, selection produces two fewer correct answers than always accepting the hybrid candidate, while still yielding two more than always retaining Phase I. Each two-answer difference corresponds to approximately $100 \times 2/69 = 2.9$ percentage points. The separately rounded accuracies need not reproduce this count-based difference exactly when subtracted. This diagnostic shows that confidence arbitration can discard correct alternatives despite its positive numerical difference on the full evaluation set. Correctness counts alone neither identify the visual evidence supporting an answer nor distinguish poor localization, loss of spatial detail, and cross-context score comparability as causes of the reversions.

5 Conclusion

Cog-Fovea treats additional observation acquisition and answer replacement as separate decisions within a frozen video reasoning model. It combines global inspection, conditional temporal resampling, and deterministic confidence arbitration. Answer-only inference reaches 60.2% accuracy on Video-MMMU at 6.3 seconds per query, with approximately 35% of queries receiving additional inspection. Arbitration yields a numerical gain of 0.6 percentage points over accepting every available routed answer on the full evaluation set, but loses correct revisions on the narrow-window subset. These findings support conditional visual inspection under the evaluated configuration and identify cross-context answer selection as a remaining limitation. Further limitations are discussed in Appendix E.

AI USE STATEMENT

OpenAI Codex was used to assist manuscript drafting, restructuring, and language editing.

ETHICS STATEMENT

This work studies inference-time video reasoning using existing benchmarks. Conditional inspection does not eliminate errors or biases inherited from the underlying model and data. Applications involving personal or surveillance video require appropriate consent, privacy safeguards, and compliance with applicable data-use restrictions. Model confidence should not be treated as a safety guarantee, and the reported benchmark results do not establish suitability for high-stakes decisions.

REPRODUCIBILITY STATEMENT

Section 3 specifies the inference stages, confidence scores, routing thresholds, frame-selection rules, fallback behavior, and computation accounting. Appendix D describes the backbone, evaluation protocols, hardware, and inference settings. Appendix A provides the inference prompts, and Appendix B documents the sampling controls and component comparisons. The assumptions underlying the information and coverage analyses are stated in Section 3.1 and Appendix C.1, respectively. Section 4.2 reports count-based answer-selection diagnostics to clarify the relationship between aggregate accuracy and individual revisions.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. Appo: Attention-guided perception policy optimization for video reasoning. *arXiv preprint arXiv:2602.23823*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025a.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Lover1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*, 2025b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. Visioncoach: Reinforcing grounded video reasoning via visual-perception prompting. *arXiv preprint arXiv:2603.14659*, 2026.

Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. Frameoracle: Learning what to see and how much to see in videos. *arXiv preprint arXiv:2510.03584*, 2025a.

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

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11369–11380, 2026a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024.

Xuchen Li, Xuzhao Li, Shiyu Hu, and Kaiqi Huang. Select less, reason more: Prioritizing evidence purity for video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25621–25632, 2026b.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

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

Chaofan Ma, Zhenjie Mao, Yuhuan Yang, Fanqin Zeng, Yue Shi, Yingjie Zhou, Xiaofeng Cao, and Jiangchao Yao. Reason, then re-reason: Cross-view revisiting improves spatial reasoning. *arXiv preprint arXiv:2606.11683*, 2026a.

Martin Q Ma, Yuxiao Qu, Aditya Agrawal, Willis Guo, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Act2see: Emergent active visual perception for video reasoning. *arXiv preprint arXiv:2605.01657*, 2026b.

Ruiying Peng, Xueyu Wu, Jing Lei, Lu Hou, Yuanzheng Ma, and Xiao-Hui Li. Deeper thought, weaker aim: Understanding and mitigating perceptual impairment during reasoning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12064–12073, 2026.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16944–16953, 2026.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6613–6629, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28114–28128, 2025.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. Season: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11096–11105, 2026.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd annual meeting of the association for computational linguistics*, pp. 133–138, 1994.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, pp. 20–37. Springer, 2024.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.