READ FIRST, LOOK ON DOUBT: A TEXT-FIRST FRAMEWORK FOR LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models have advanced long-video question answering chiefly by looking harder, through longer frame contexts, compressed visual tokens and multimodal memory agents, all treating the collapse of visual coverage as the problem to solve. We report two observations that point the other way. As videos lengthen, the language of the footage, its speech, narration and on-screen text, answers a growing share of what is asked on its own, and accuracy from the language track alone grows steadily stronger with duration while accuracy from visual input collapses. Visual information added on top of the text in turn loses its marginal value, which falls from $+14.6$ points on the shortest clips to -13.7 at the longest tier. In response, we propose READFIRST, a text-first framework that inverts the standard allocation. It distils every channel of a video into a multi-scale textual memory offline and answers by reading at query time, escalating to a targeted visual inspection only when two derivations of the same answer disagree, thereby spending perception only where the text runs out. Extensive experiments demonstrate that READFIRST achieves state-of-the-art performance, reaching an average accuracy of 74.1 across five long-video benchmarks while resolving over 90% of questions in text alone.

1 INTRODUCTION

Video question answering has outgrown the clip. Benchmarks have advanced from seconds-long segments to hour-scale footage and multi-day egocentric logs (Wang et al., 2025a; Chandrasegaran et al., 2024; Yang et al., 2025a), and the field has met that growth by making models look harder. Frontier systems such as GPT-5 (OpenAI, 2025), Gemini (Comanici et al., 2025) and Qwen3-VL (Bai et al., 2025) ingest frames directly; open systems push more footage through the visual pathway with hierarchical token compression (Li et al., 2026), spatiotemporal downsampling (Shen et al., 2025a) and fixed-size visual memories (Song et al., 2024; He et al., 2024); and reinforcement learning sharpens temporal grounding (Wang et al., 2025c;d). Agentic systems carry the same commitment into memory, culminating in engines that maintain episodic, semantic and visual banks with GPU embedding retrieval (Lin et al., 2026; Long et al., 2026; Tian et al., 2026; Yeo et al., 2026). Every one of these designs spends its query-time budget on perception, an allocation that suits minute-scale video, where a frame budget covers the whole clip. However, as footage stretches toward hours, that allocation meets a structural obstacle.

The obstacle originates from simple arithmetic. A fixed evaluation budget shows a model a few dozen frames regardless of length, so the budget that covers every frame of a one-minute clip covers only a vanishing fraction of an hour, and watching degrades into blind sampling. This collapse is well recognised. The community has responded by compressing more footage into the budget (Li et al., 2026; Shen et al., 2025a), selecting frames more cleverly (Tang et al., 2025; Wang et al., 2024; Xie et al., 2026), and consolidating past observations into memory (Song et al., 2024; Lin et al., 2026; Yeo et al., 2026). All three keep the pixels in the query path, whether inside the frame budget or behind a store the question still reaches into. In this work, we report two observations suggesting the budget points at the wrong channel. First, as videos lengthen, the language of the footage, its speech, narration and on-screen text, answers a growing share of what is asked on its own. With the model and prompt held fixed, accuracy from the language track alone climbs steadily with duration,

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.

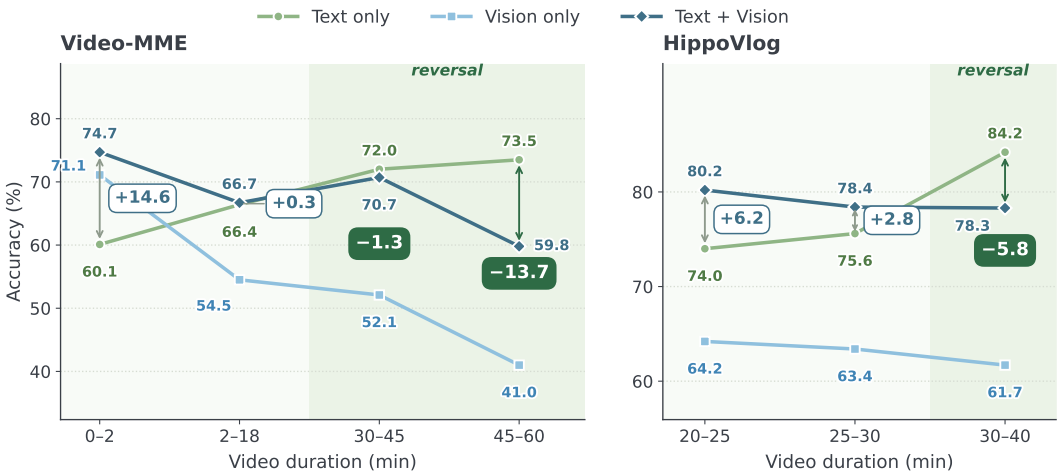

Figure 1: **The marginal value of looking decays with duration.** Accuracy of MiniMax-M3 given the language track alone (*Text only*), visual input alone (*Vision only*), or both (*Text + Vision*); boxed numbers give the per-tier gap Text+Vision − Text. On Video-MME (left), text climbs $60.1 \rightarrow 73.5$ while frames collapse $71.1 \rightarrow 41.0$, and the gain from adding frames shrinks from $+14.6$ to -13.7; on HippoVlog (right), the same decay runs $+6.2 \rightarrow -5.8$. The value of added visual input decays monotonically with duration and turns negative on both benchmarks. Protocol and per-tier statistics in Appendix A.

while accuracy from frames collapses. Second, frames added on top of the language track stop helping as duration grows, and their marginal value turns negative at the longest durations.

To examine both observations, we conduct a controlled modality ablation. The same model answers the same questions from the language track alone, from sampled frames alone, or from both together, swept across the full duration range of Video-MME and replicated on HippoVlog. The results, summarised in Figure 1, confirm both observations. Text-only accuracy climbs with duration while frames-only collapses, and the gain from adding frames to text shrinks tier by tier, from $+14.6$ points on the shortest clips to below zero at the longest tier of each benchmark.

The arithmetic that breaks a query-time budget does not bind an offline one. Perception run before any question arrives covers the whole recording at a steady rate, and what it writes down is text, the medium these measurements show holding up as footage lengthens. READFIRST is built on that reading. Following caption- and memory-based systems (Zhang et al., 2024a; Wang et al., 2025e; Lin et al., 2026; Yeo et al., 2026), it separates an offline build from online inference; the build distils every channel of the video into a multi-scale textual memory, fusing at the finest scale what is seen and what is said within the same window. The memory reads no question, no option and no label, so one pass serves every question that follows, and at query time a reading agent sees that memory and nothing else.

Some answers still live only in pixels, a colour, a spatial layout, an action nobody narrates, and where the language track is missing entirely, frames beat text by a decisive margin. A uniform policy fails at both ends. Withhold frames from every question and those go unanswered; attach frames to every question and the marginal value of the addition has already turned negative on long footage. The right unit of allocation is the question.

Humans facing the same predicament rarely re-watch hours of footage. They consult notes, transcripts and recollections first, and return to the screen only when their recollections conflict. A learned estimator or a tuned threshold could make the same call, at the cost of a calibration set for every benchmark and every reasoning model it runs on. READFIRST takes disagreement as the signal instead: before committing, it re-runs the same frozen answering function on identical evidence, and disagreement between the two derivations marks the questions the memory cannot decide. Only those escalate to a targeted visual inspection.

Extensive experiments demonstrate that READFIRST achieves state-of-the-art performance across five long-video benchmarks, reaching an average accuracy of 74.1 and surpassing the strongest multimodal memory system under the same reasoning model. The margin does not rest on looking. Reading alone already exceeds that system, and over 90% of questions never leave the text path. The framework holds under a weaker open-tier reasoning model, placing above every system that watches frames directly. Ablation studies further validate the effectiveness of the multi-scale textual memory, the reading loop and the second-opinion gate.

We summarise our contributions as follows:

1. We show through controlled modality ablations that a text record answers most long-video questions on its own while the marginal value of frames added at query time decays with duration and turns negative, recasting long-video question answering as a choice of medium and a per-question allocation.
2. We propose READFIRST, a text-first framework that transcribes perception into a question-independent multi-scale textual memory offline, fusing sight and speech at the finest scale, answers by reading alone, and grants pixel access to the doubting minority through a training-free second-opinion gate.
3. We conduct extensive experiments on five long-video benchmarks, where READFIRST achieves state-of-the-art performance, and ablation studies validate the effectiveness of each component.

2 RELATED WORK

Long video understanding. Work on longer video runs along four routes. One extends the native context so that more frames fit a single pass (Zhang et al., 2024b; 2025; Bai et al., 2025). A second compresses what enters or stays in the model, through hierarchical token compression (Li et al., 2026), spatiotemporal downsampling (Shen et al., 2025a; Hwang et al., 2024), consolidation across hybrid or continuous time (Wang et al., 2025b; Santos et al., 2025), and streaming schemes that hold hour-long footage in a bounded visual memory (Chen et al., 2026c; Song et al., 2026; Sun et al., 2026). A third selects where a fixed budget looks, from adaptive keyframe sampling (Tang et al., 2025) to inference-time active perception (Ma et al., 2025). A fourth sharpens temporal reasoning through post-training (Wang et al., 2025c;d). Every route expands what a model can watch per query, yet the budget still covers a vanishing fraction of hour-long footage. We target the allocation instead, moving perception out of the query path rather than scaling what it sees.

Memory-based video agents. A second family answers by search, pairing a language model with tools or memory. Frame-seeking agents iteratively locate the windows a question needs (Wang et al., 2024; Fan et al., 2024; Xie et al., 2026; Ouyang et al., 2025; Zhang et al., 2026), and later designs add exploration-exploitation policies, hypothesis verification, interactive perception and programmatic probing (Yang et al., 2025b; Wang et al., 2026; Li et al., 2025; Rauf et al., 2026). On the memory side, tool-chained egocentric reasoners (Tian et al., 2026), hippocampus-inspired and lifelong stores (Lin et al., 2026; Long et al., 2026; Kim et al., 2026; Chen et al., 2026b), and hierarchical or graph-structured memories with agentic retrieval (Yin et al., 2026; Chen et al., 2026a) extend the family. WorldMM (Yeo et al., 2026) culminates the line, maintaining episodic, semantic and visual banks with GPU embedding retrieval (Meng et al., 2026). Despite this sophistication, the family never commits to an abstraction, building every store over the full footage and keeping the frames reachable at query time. WorldMM makes the fallback selective, its retrieval agent picking one bank per round and its paper noting that always pairing captions with frames can distract the model, yet the choice still ranges over banks that hold the full visual signal, behind a query encoder that every question pays for on GPU. In contrast, we distil visual detail into text once and pay for pixels only on measured doubt, which outperforms WorldMM under its own protocol and reasoning model while removing the GPU from the query path entirely.

Language-based video understanding. A third line converts video to language and reasons over the text, through Socratic composition (Zeng et al., 2023), dense-caption pipelines (Zhang et al., 2024a), query-adaptive caption hierarchies (Wang et al., 2025e), and sparse-memory token schemes (Song et al., 2024; He et al., 2024). Retrieval-augmented systems extend the line, fetching evidence

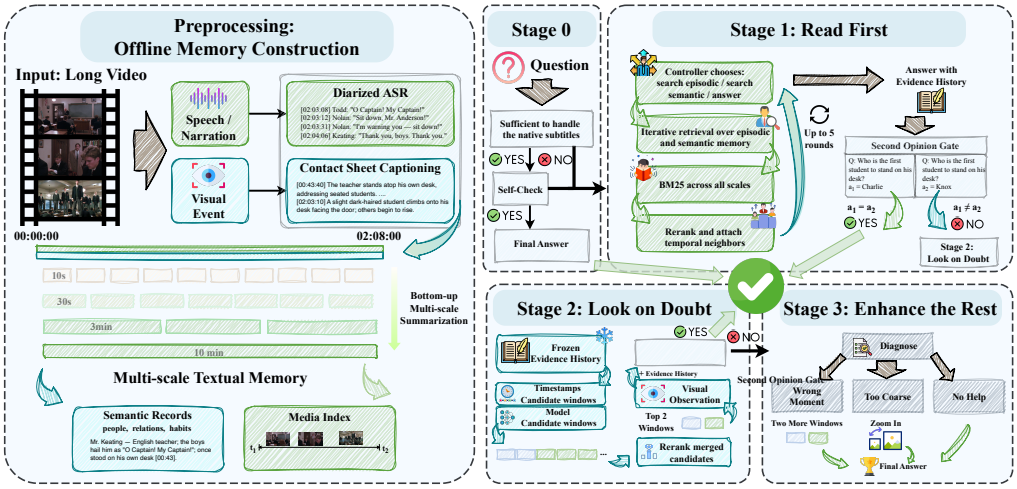

Figure 2: **The READFIRST framework.** Offline, one pass per video distils diarized speech and captioned sight into a multi-scale textual memory, alongside semantic records and a media index that only the gated stages read. Online, Stage 0 answers from a transcript that fits the budget whole; Stage 1 (*Read First*) retrieves and answers from the memory alone; the second-opinion gate re-derives the answer from identical evidence and commits on agreement; disagreement enters Stage 2 (*Look on Doubt*), a bounded verdict-free inspection, and any residual split enters Stage 3 (*Enhance the Rest*), which fetches once more and commits unconditionally.

per question with neural encoders (Guo et al., 2025; Gutiérrez et al., 2024; Luo et al., 2025; Jeong et al., 2025; Xu et al., 2025; Xue et al., 2025; Shen et al., 2025b). Yet a recent audit literature documents how much of the language channel this stack leaves unused. Benchmarks never required listening, so systems ship without speech encoders (Kim & Seo, 2026); audio pathways prove vestigial once text shortcuts disappear (Yang et al., 2025c); and subtitles rank among the densest available signals yet go routinely ignored (Fang et al., 2024). These audits diagnose the neglect. We close it from the systems side, distilling diarized speech and sight into one multi-scale textual memory served by plain lexical retrieval (Robertson et al., 1994), and the resulting system outperforms every retrieval-augmented system above.

3 READFIRST: READING FIRST, LOOKING ON DOUBT

3.1 OVERVIEW

READFIRST answers questions about long video by reading first and looking only on doubt; Figure 2 overviews the pipeline. An offline pass runs once per video and distils every channel, speech, visual events and on-screen text, into a multi-scale textual memory (§3.2). At query time, a question whose video carries a transcript compact enough to read whole resolves in one pass over it (Stage 0, §3.3). Every other question enters Stage 1, *Read First*, a reading loop that gathers evidence from the memory through iterative retrieval (§3.4). Before the loop's answer commits, a second-opinion gate re-derives it from identical evidence (§3.5). Questions the gate flags enter Stage 2, *Look on Doubt*, a targeted inspection on frozen evidence (§3.6), and the residue on which two fused derivations still disagree enters Stage 3, *Enhance the Rest*, which diagnoses the conflict, fetches once more, and commits the final derivation unconditionally (§3.7).

3.2 OFFLINE MEMORY CONSTRUCTION

Questions about a video arrive many at a time while the footage itself never changes, so we perform perception once, offline, and share the result across every question that follows:

$$\mathcal{M} = \bigcup_{s \in \mathcal{S}} \Phi_s(V), \qquad \mathcal{S} \subseteq \{10\,\text{s}, 30\,\text{s}, 3\,\text{min}, 10\,\text{min}, 1\,\text{h}\}, \tag{1}$$

where Φ_s converts each window of video V at scale s into a text record. Three record types compose the memory.

Episodic records. At the finest scale, Φ captions every 10-second segment from a 1 fps contact sheet and merges what is visible with what is spoken in the same window. A locally run diarized ASR model (MOSI.AI, 2026) transcribes the speech, and its acoustic alignment attaches each utterance to the window in which it was spoken, so that a single record describes one moment across sight and sound. Each record names the visual events, the words, the entities present and any on-screen text. Coarser scales then summarise bottom-up, each summary citing the finer records it distils, which keeps every scale traceable to the footage below it and gives retrieval a granularity to match whatever span a question asks about. Hour-scale video runs this ladder from 10 s to 10 min, and week-long egocentric video from 30 s to 1 h.

Semantic records. Statements about people, relations and habits carry the knowledge that recurs across a recording rather than living at one timestamp. The egocentric sets ship with a pre-built semantic bank; elsewhere, a single text call derives these statements at query time from episodic records the reader has already cited, so semantic memory grows out of used evidence rather than pre-computed speculation.

The media index. A final index maps any time range back to its frames. Only the gated stages read it, so pixels stay reachable without ever entering the text path.

3.3 Stage 0: reading whole before retrieving

Retrieval exists to keep evidence dense when the text outgrows the context, and applying it to a text that fits can only lose information. An iterative loop reads a transcript as retrieved fragments, and what a single pass would surface can vanish between them. Therefore, when the video behind a question carries a transcript within a budget of 32k tokens, ReadFirst skips the loops entirely. A frozen answering function F, shared by every stage that follows, reads the flat, timestamped transcript in one pass and answers, and a brief self-check reads the same transcript and judges whether it settles the question. Questions the check declines, together with those whose transcript is missing or over budget, enter the reading loop.

3.4 Stage 1: Read First

For the questions that remain, the memory is far too large to read whole, so the reading loop assembles the evidence each question needs over at most five retrieval rounds. In each round, a controller decides whether to search the episodic bank, to search the semantic bank, or to answer; the action space holds nothing else, so the reader can only read. The loop opens by retrieving with the question itself, and answering becomes available from the second round.

A search runs against every scale of the memory at once. The query expands with simple word-form variants, and full-bank BM25 on CPU (Robertson et al., 1994) scores every record. Recall keeps a fixed candidate budget per scale, so that a one-line hour summary survives alongside thousands of 10-second captions, and a single answer-free call reranks the candidates, seeing each record's scale, span and text but never the question's options, and free to select nothing. Selected records attach their immediate temporal neighbours, restoring the context around events that cross window boundaries; questions that cite an explicit playback time anchor retrieval to that moment; and a failed call enters the history as an explicit error record, so the loop continues rather than silently retrying. The whole path holds no embedding model, no vector database and no GPU.

Writing the controller as π, the query it issues in round k as u_k, and the retriever as $\rho_{\mathcal{M}}$, the loop compounds an evidence history h for the question q:

$$u_k = \pi(q,\, h_{k-1}), \qquad h_k = h_{k-1} \oplus \big(u_k,\ \rho_{\mathcal{M}}(u_k)\big), \tag{2}$$

and when the loop ends, because the controller elects to answer or the round budget is spent, the frozen function F of §3.3 reads the final history and produces the reader's answer $a_1 = F(h)$.

3.5 THE SECOND-OPINION GATE

The reading loop ends with an answer, and the system must then decide, per question, whether that answer warrants a look. Learned confidence or a tuned threshold could make this decision, but both must recalibrate whenever the benchmark or the reasoning model changes. We instead derive the answer a second time from the same evidence, $a_2 \sim F(h)$, and escalate only on disagreement:

$$\text{FIRE} \;=\; \mathbb{1}[\, a_1 \neq a_2 \,]\,. \tag{3}$$

Two derivations disagreeing over the same evidence marks a question that evidence leaves open, and an open decision is the one condition under which further evidence can change the outcome. Three properties follow, measured in §4.3 and Appendix C. The gate is calibration-free, since it compares two answers for equality and interprets no scalar. It is adaptive, since firing tracks whatever makes this model unstable on this data. And it is model-conditional, since a weaker model fires more often. Questions that pass commit a_1 unchanged, and only the disagreeing minority proceeds to visual inspection.

3.6 STAGE 2: LOOK ON DOUBT

When the gate fires, Stage 2 adds perception while leaving untouched the evidence that triggered it, through three commitments.

Frozen evidence. The loop inherits the reader's evidence history wholesale and issues no new memory searches, because adding perception to fixed evidence answers one question cleanly, whether the text was sufficient, while re-rolling retrieval would confound that answer with a second draw.

Blinded targeting. Candidate time windows arrive from two channels, a code channel that reads timestamps off the inherited evidence and a model channel that proposes complementary ranges from a coarse overview of the bank. The merged candidates lose their origin labels before a single verification call ranks them, so a model-proposed range earns no deference over a code-derived one, and the top two windows $w^{\star}_{1:2}$ render as eight-frame contact sheets.

Verdict-free inspection. The inspection informs the final decision rather than making it, so each inspected window yields a bounded textual observation of what is visible instead of a verdict:

$$o \;=\; \Omega(w^{\star}_{1:2})\,, \qquad a \;=\; F(h \oplus o)\,, \tag{4}$$

where Ω describes the rendered windows and its observations append to the frozen history as text, so that F re-derives the answer from text and observation together rather than from a glimpse of pixels alone. Two independent derivations of F over the extended history then decide the commit. Agreement ends the question, and the residue on which the two still disagree enters Stage 3.

3.7 STAGE 3: ENHANCE THE REST

Disagreement that survives inspection means the vision already gathered did not settle the question, and the useful next step depends on why. A single diagnosis call therefore classifies the residual conflict into one of three causes: the inspected windows sit at the wrong moment, the inspection hit the right moment but too coarsely, or no further vision is likely to help. Each cause routes to its own fetch. The first fetches up to two new windows from the candidates the earlier cap set aside; the second zooms into the closest inspected window, magnifying its frames spatially; the third fetches nothing visual. Every route also widens the local transcript around the target window. A last derivation of F over everything gathered then commits as the final answer, unconditionally.

4 EXPERIMENTS

4.1 SETUP

Benchmarks. We evaluate on five long-video suites. EgoLife (Yang et al., 2025a) records a shared household for a full week through head-mounted cameras and asks about events days apart. Ego-R1 (Tian et al., 2026) poses multi-step reasoning over week-scale egocentric footage. HippoVlog (Lin et al., 2026) collects vlogs of everyday activities and travel, most twenty to forty minutes long.

Table 1: **Main results on five long-video benchmarks.** We compare READFIRST against fifteen strong published systems (Yeo et al., 2026) spanning direct frame access, retrieval augmentation and multimodal memory. Average is the mean over the five benchmarks, and **bold** marks the best per column. READFIRST sets the state of the art with the best average accuracy.

System	EgoLife	Video-MME (L)	Ego-R1	HippoVlog	LVBench	Average
Video LLMs (direct)						
Qwen3-VL-8B (Bai et al., 2025)	38.6	61.0	35.7	74.4	48.3	51.6
Gemini 2.5 Pro (Comanici et al., 2025)	46.4	55.7	46.7	72.0	57.0	55.6
GPT-5 (OpenAI, 2025)	48.6	74.3	46.3	75.7	60.4	61.1
VideoChat-Flash (Li et al., 2026)	34.2	44.1	42.7	58.0	33.2	42.4
Time-R1 (Wang et al., 2025c)	48.8	37.6	48.0	54.6	31.1	44.0
Video-RTS (Wang et al., 2025d)	48.2	47.9	47.4	59.0	39.8	48.6
Retrieval-augmented						
LightRAG (Guo et al., 2025)	48.8	46.6	52.3	47.4	30.4	45.1
HippoRAG (Gutiérrez et al., 2024)	59.6	52.1	56.0	63.2	54.0	57.0
Video-RAG (Luo et al., 2025)	55.4	55.4	49.7	65.1	33.1	51.7
EgoRAG (Yang et al., 2025a)	52.0	41.1	49.0	57.5	32.2	46.4
Memory & agent systems						
Ego-R1 (Tian et al., 2026)	53.0	42.7	52.0	58.8	34.1	48.1
HippoMM (Lin et al., 2026)	54.6	41.6	53.0	71.9	38.2	51.8
M3-Agent (Long et al., 2026)	53.5	55.3	52.0	65.5	49.3	55.1
WorldMM-8B (Yeo et al., 2026)	56.4	66.0	52.0	69.7	55.4	59.9
WorldMM-GPT (Yeo et al., 2026)	65.6	76.6	65.3	78.3	61.9	69.5
Ours (text-first)						
READFIRST — MiniMax (ours)	55.6	73.4	69.0	80.0	62.8	68.2
READFIRST — GPT-5 (ours)	**67.4**	**79.9**	**72.7**	**81.1**	**69.6**	**74.1**

LVBench (Wang et al., 2025a) spans hour-long web videos across diverse genres, and Video-MME Long (Fu et al., 2025) holds the 30-to-60 minute split of a general video benchmark, with native subtitles. Together they pose 4,194 four-way multiple-choice questions.

Models. READFIRST runs with GPT-5 as its main reasoning model, the same model that powers the strongest published baseline, WorldMM-GPT, so the headline comparison holds the reasoning model fixed. A second configuration substitutes the open-tier MiniMax-M3 (Lai et al., 2026) with everything else unchanged, testing whether the framework survives a weaker reasoning model.

Baselines. We benchmark against fifteen strong published systems spanning three paradigms: frontier and open video LLMs with direct frame access, led by GPT-5 and Gemini (OpenAI, 2025; Comanici et al., 2025; Bai et al., 2025; Li et al., 2026; Wang et al., 2025c;d); retrieval-augmented systems (Guo et al., 2025; Gutiérrez et al., 2024; Luo et al., 2025; Yang et al., 2025a); and multimodal memory agents (Tian et al., 2026; Lin et al., 2026; Long et al., 2026), culminating in WorldMM (Yeo et al., 2026), the state of the art on these suites.

4.2 MAIN RESULTS

From the evaluation across the five benchmarks, presented in Table 1, we draw three findings.

Limits of perception-first baselines. Across the three baseline paradigms, accuracy tracks the evidence that reaches the reasoning model rather than the model's capacity. GPT-5 with direct frame access reaches 61.1 on a budget that covers a vanishing fraction of hour-long footage, and WorldMM-GPT reaches 69.5, the strongest published result, by retrieving evidence per question while still keeping perception inside the query loop. The binding constraint on these suites is therefore the evidence assembled for reasoning.

Effectiveness and generality of READFIRST. Our method achieves state-of-the-art performance, reaching an average accuracy of **74.1**, the best among all seventeen systems, surpassing the strongest multimodal memory system by $+4.6$ under the same reasoning model and winning all five datasets,

Table 2: **Reading versus the full system.** Read is the reader's answer before any visual stage, Full is the deployed system with gated escalation, and Resolved is the share of questions answered without decoding a single pixel, the transcript-branch acceptances plus the questions whose gate never fires, on identical questions under both reasoning models. Reading alone already clears the strongest multimodal baseline with GPT-5, and both models resolve most questions without any visual stage.

	GPT-5			MiniMax-M3		
Dataset	Read	Full	Resolved	Read	Full	Resolved
EgoLife	64.40	67.40	90.0%	53.20	55.60	70.2%
Video-MME (L)	79.44	79.89	96.9%	72.89	73.44	93.6%
Ego-R1	71.67	72.67	91.7%	66.67	69.00	81.3%
HippoVlog	80.31	81.15	96.5%	78.44	80.00	93.2%
LVBench	68.58	69.62	93.5%	61.54	62.84	84.3%
Average	**72.88**	**74.15**	**93.7%**	**66.55**	**68.18**	**84.5%**

Table 3: **Component ablation.** We report the impact of incrementally adding each component, starting from the raw transcript in a single call, with parenthesised values giving the gain over the preceding rung. Accuracy climbs at every rung, confirming that each component contributes to the final margin.

Configuration	Accuracy
Transcript in context	52.4
+ multi-scale memory	57.4 (+5.0)
+ distilled captions	64.4 (+7.0)
+ gated escalation	**67.4** (+3.0)

Table 4: **Escalation trigger ablation.** Comparison between a confidence-based gate, which escalates when the model reports low certainty, and the second-opinion gate. The second opinion answers more accurately while firing less often.

Trigger	Accuracy	Fire rate
Confidence	64.6	12.8%
Second opinion	**67.4** (+2.8)	10.0%

with margins running from +1.8 on EgoLife to +7.7 on LVBench. Notably, with the open-tier MiniMax-M3 substituted in, READFIRST still reaches 68.2, above every system with direct frame access, which attests that the framework rather than the reasoning model carries the result.

Reading carries the result. We separate reading from looking on the same 4,194 questions in Table 2. Before any visual stage runs, the reader scores an average of **72.9**, +11.8 over the same model watching the video directly and already above the strongest multimodal baseline, and on average 93.7% of questions resolve without decoding a single pixel, never below 90% on any single benchmark. Under MiniMax-M3 the division of labour is the same, reading resolving 84.5% of them on average, so the division is a property of the framework rather than of one model. The weaker model's gate also fires more than twice as often, and the two models' escalated sets overlap 2.39× beyond chance, which reads as doubt tracking difficulty conditioned on the model that reads: the same rule grants a weaker reader more looking. §4.3 quantifies what that looking adds.

4.3 ABLATION STUDIES AND ANALYSIS

To validate the mechanisms underpinning READFIRST, we ablate the three decisions the design makes, namely what goes into memory, how long the reader searches it, and when to escalate. Unless stated otherwise, ablations run with GPT-5 on EgoLife's 500 frozen questions under the main configuration.

Component ablation. Table 3 validates that the margin stems from the framework's components rather than from the reasoning model, by rebuilding the system one component at a time and holding the model, the questions and the policy fixed. The bare model reading the raw transcript in a single call reaches 52.4. Restructuring the same speech into the multi-scale memory raises accuracy to 57.4 (+5.0), as retrieval over scale-matched records replaces one diluted context. Adding the captions

distilled from frames raises it to 64.4 (+7.0), giving the reader sight it never pays for at query time, and attaching the gated escalation completes the system at 67.4 (+3.0), recovering questions the text leaves open. Each component contributes a measured gain on top of the ones before it, validating that every design in the framework is both purposeful and effective.

Reading-loop budget. Figure 3 asks where the retrieval budget should sit by deploying the reader independently at one, three, five, seven and ten rounds. Accuracy climbs steeply up to the deployed five, +8.4 from one round to five, then nearly stops, with the two budgets beyond adding +1.2 between them while the retrieval spent per question keeps growing. The deployed budget therefore balances accuracy against cost: below five the reader leaves accuracy on the table, and beyond five every additional round buys little at full price.

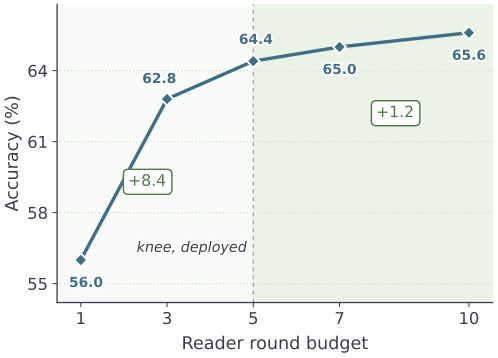

Figure 3: **Reading-loop budget.** Accuracy climbs steeply up to the deployed five rounds, +8.4 points from one round to five, and barely beyond, +1.2 across seven and ten, while the retrieval spent per question keeps growing; five rounds balances accuracy against cost.

Escalation trigger. Table 4 validates the disagreement trigger against the standard alternative, a confidence-based gate that asks the model how sure it is and escalates below a threshold. The second opinion reaches 67.4 against 64.6 for confidence, a +2.8 margin, while firing on fewer questions, 10.0% against 12.8%. Comparing two derivations for equality therefore selects escalation targets better than a verbalised scalar, with nothing to calibrate.

What the visual stages add. Table 5 prices the two visual stages on top of reading across the five benchmarks. The gated look contributes +0.98 points of average accuracy and Stage 3 a further +0.29. The aggregate understates the effect, because escalation only touches the questions the gate selects: on the 237 escalated questions the reader scores 34.2, and one targeted look lifts them to 51.1. Part of that lift is selection rather than perception, since the look both adds observations and re-derives the answer, and Appendix C bounds the two shares with controls the trace already records.

Table 5: **What the visual stages add.** Accuracy as each visual stage joins the reader. The All column averages the five benchmarks, and the Escalated column pools the 237 questions the gate escalates.

Configuration	All	Escalated
Reading alone	72.88	34.2
+ Stage 2	73.86	51.1
+ Stage 3	**74.15**	**53.6**

5 Conclusion

This work identifies a misallocation in long-video question answering. Measured at scale, answering from the language track alone strengthens with duration while answering from visual input collapses, and the marginal value of visual input added to the text turns negative. We therefore propose ReadFirst, a text-first framework that distils perception offline into a multi-scale textual memory, answers by reading, and escalates to perception only when two derivations of the same answer disagree. Across five long-video benchmarks, reading alone surpasses the strongest multimodal memory system, doubt-gated looking sets the state of the art, and ablations validate each component. We hope this allocation paves the way toward systems that read first and look only on doubt.

AI USE STATEMENT

We used generative AI tools to polish the language of this paper, to search for and summarise related literature, and to assist with writing the implementation and analysis code. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

REFERENCES

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

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, et al. HourVideo: 1-hour Video-Language understanding. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Cong Chen, Guo Gan, Kaixiang Ji, ZhaoYang Zhang, Zhen Yang, Guangming Yao, Hao Chen, Jingdong Chen, et al. MemDreamer: Decoupling perception and reasoning for long video understanding via hierarchical graph memory and agentic retrieval mechanism. *arXiv preprint arXiv:2606.07512*, 2026a.

Guo Chen, Lidong Lu, Yicheng Liu, Liangrui Dong, Lidong Zou, Jixin Lv, Zhenquan Li, Xinyi Mao, et al. Towards multimodal lifelong understanding: A dataset and agentic baseline. *arXiv preprint arXiv:2603.05484*, 2026b.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. StreamingTOM: Streaming token compression for efficient video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026c.

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

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

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. MMBench-Video: A Long-Form Multi-Shot benchmark for holistic video understanding. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, et al. Video-MME: The First-Ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast Retrieval-Augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired Long-Term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. EVEREST: Efficient masked video autoencoder by removing redundant spatiotemporal tokens. In *International Conference on Machine Learning (ICML)*, 2024.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-Augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Geewook Kim and Minjoon Seo. Do modern Video-LLMs need to listen? a benchmark audit and scalable remedy. In *Interspeech*, 2026.

Kangsan Kim, Yanlai Yang, Suji Kim, Woongyeong Yeo, Youngwan Lee, Mengye Ren, and Sung Ju Hwang. MA-EgoQA: Question answering over egocentric videos from multiple embodied agents. *arXiv preprint arXiv:2603.09827*, 2026.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, et al. MiniMax sparse attention, 2026.

Jiahua Li, Zhanhe Zhang, Chenghao Xu, Zhe Xu, Kun Wei, Xu Yang, and Cheng Deng. Perceive, verify and understand long video: Multi-Granular perception and active verification via interactive agents. *arXiv preprint arXiv:2509.24943*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, et al. VideoChat-Flash: Hierarchical compression for Long-Context video modeling. In *International Conference on Learning Representations (ICLR)*, 2026.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai "Helen" Li, and Yiran Chen. HippoMM: Hippocampal-inspired multimodal memory for long audiovisual event understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Findings Track*, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with Long-Term memory. In *International Conference on Learning Representations (ICLR)*, 2026.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, et al. Video-RAG: Visually-aligned Retrieval-Augmented long video comprehension. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Martin Q. Ma, Willis Guo, Aditya Agrawal, Ankit Gupta, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Video active perception: Effective Inference-Time Long-Form video understanding with Vision-Language models. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, et al. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*, 2026.

MOSI.AI. MOSS transcribe diarize technical report, 2026.

OpenAI. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over Multi-Scale visual evidence. *arXiv preprint arXiv:2510.20470*, 2025.

Awais Rauf, Ahmed Hasssan, and Greg Slabaugh. HPP: Hierarchical programmatic probing for long video understanding by decoupling perception and reasoning. *arXiv preprint arXiv:2606.21734*, 2026.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1994.

Saul Santos, António Farinhas, Daniel C. McNamee, and André F. T. Martins. ∞-video: A Training-Free approach to long video understanding via Continuous-Time memory consolidation. In *International Conference on Machine Learning (ICML)*, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, et al. LongVU: Spatiotemporal adaptive compression for long Video-Language understanding. In *International Conference on Machine Learning (ICML)*, 2025a.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based Retrieval-Reasoning-Augmented generation for long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Baiyang Song, Yuli Lin, Qiong Wu, Tao Chen, Jun Peng, Xiao Chen, Yiyi Zhou, and Rongrong Ji. Towards a dynamic and fixed-budget memory bank for efficient streaming video understanding. *arXiv preprint arXiv:2606.25658*, 2026.

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

Guangzhi Sun, Yixuan Li, Yudong Yang, and Chao Zhang. OmniMem: Perturbation-aware memory compression for streaming audio-visual LLMs. *arXiv preprint arXiv:2606.07577*, 2026.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, et al. Ego-R1: Chain-of-Tool-Thought for Ultra-Long egocentric video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, et al. LVBench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

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 (ECCV)*, 2024.

Xidong Wang, Dingjie Song, Shunian Chen, Junyin Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang. LongLLaVA: Scaling multi-modal LLMs to 1000 images efficiently via a hybrid architecture. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, et al. Time-R1: Post-Training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *Empirical Methods in Natural Language Processing (EMNLP)*, 2025d.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025e.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced Multi-Turn reasoning for long video understanding. In *International Conference on Machine Learning (ICML)*, 2026.

Zeyu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. E-VRAG: Enhancing long video understanding with Resource-Efficient retrieval augmented generation. *arXiv preprint arXiv:2508.01546*, 2025.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. AdaVideoRAG: Omni-Contextual adaptive Retrieval-Augmented efficient long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, et al. EgoLife: Towards egocentric life assistant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Te Yang, Xiangyu Zhu, Bo Wang, Quan Chen, Peng Jiang, and Zhen Lei. EEA: Exploration-exploitation agent for long video understanding. *arXiv preprint arXiv:2512.03500*, 2025b.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, et al. Audio-centric video understanding benchmark without text shortcut. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2025c.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, et al. Socratic models: Composing Zero-Shot multimodal reasoning with language. In *International Conference on Learning Representations (ICLR)*, 2023.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for Long-Range video Question-Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024a.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. ReWatch-R1: Boosting complex video reasoning in large Vision-Language models through agentic data synthesis. In *International Conference on Learning Representations (ICLR)*, 2026.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, et al. InternLM-XComposer-2.5: A versatile large vision language model supporting Long-Contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025.

A THE MODALITY ABLATION: PROTOCOL AND STATISTICS

Protocol. The three cells share one prompt, the official Video-MME evaluation template copied verbatim, subtitle block, question, options and the closing "The best answer is:", answered by MiniMax-M3 at temperature 0, taking the first A–D letter in the reply. The vision cells sample 16 frames uniformly at $t_i = (i + 0.5)/N \cdot D$, resized to 448 pixels on the long side at JPEG quality 80 and passed inline; the subtitle block enters untruncated, and the longest realised prompt holds 15.6k tokens. Every question counts toward accuracy, and a cell that returns no valid answer counts as wrong.

Repairs and reversals. The marginal gain in Figure 1 is the net of two flows, questions the text answered wrongly that become correct once frames are added, and questions it answered correctly that become wrong. Measured as rates, the two behave differently as duration grows. On Video-MME the share of text errors that added frames repair falls from 60.5% on the shortest clips through 39.6% and 34.1% to 18.2% at 45 to 60 minutes, while the share of correct text answers that change holds between 15.1% and 25.1% at every tier. HippoVlog repeats the pattern over its narrower range, repairs falling from 47.2% to 26.3% against reversals of 8.2% to 11.9%. The text is also right more often as videos lengthen, so the reversal share applies to a growing pool of correct answers while a shrinking repair share applies to a shrinking pool of errors, and the net crosses zero.

Language-track provenance. Video-MME releases official subtitles for part of its collection, and its remaining videos carry third-party ASR transcripts. The other four suites use transcripts from the diarized model of §3.2 throughout; external speech models transcribe and do nothing else, and answering, retrieval and selection run on the reasoning models of §4.1. The dense visual captions of the hour-scale suites are distilled from frames by MiniMax-M3, and the egocentric suites build on the EgoLifeCap captions released with EgoLife.

Frame selection. Uniform sampling is one way to spend a frame budget, and two further arms spend the same sixteen frames on selected windows, so that the vision cells of Figure 1 carry the strongest configuration we found rather than the simplest. The self-directed arm skims eight uniform frames, lets the model nominate up to five windows from what it saw, and spends its remaining eight inside them; the text-directed arm nominates windows from the subtitles, spends all sixteen inside them, and keeps the subtitles present at answer time.

Table 6 reports both against uniform sampling. Self-direction trails at every tier but one. Text-direction matches uniform sampling up to eighteen minutes, then loses 3.7 points from thirty to forty-five minutes and gains 3.0 from forty-five to sixty. Uniform sampling remains the stronger configuration on both channels, and it is the configuration Figure 1 reports.

Table 6: **Frame selection.** Accuracy of two arms that select which windows a fixed sixteen-frame budget goes to, against uniform sampling. The self-directed arm skims eight frames, nominates windows and spends its remaining eight inside them; the text-directed arm nominates windows from the subtitles and spends all sixteen inside them. Each pair is scored on the questions both of its arms answered, 1,759 for frames only and 1,854 for text plus vision, so the uniform columns here stand on smaller sets than Figure 1. Uniform sampling is the stronger configuration on both channels.

	Frames only		Text + Vision	
Duration (min)	Uniform	Self-directed	Uniform	Text-directed
0–2	70.7	60.9	72.4	72.1
2–18	55.3	51.7	66.8	66.8
30–45	53.0	50.2	70.9	67.2
45–60	43.1	44.5	60.4	63.4
Pooled	**57.1**	52.7	**68.4**	67.6

B Evaluation Details

Question sets. The five suites pose 500 questions on EgoLife, 900 on Video-MME Long, 300 on Ego-R1, 960 on HippoVlog and 1,534 on LVBench, 4,194 in total. HippoVlog publishes 1,000 questions over 25 videos at forty each; one video has been withdrawn from its public source, so its forty questions, ten from each of the benchmark's four categories, cannot be evaluated, and the remaining 24 videos are evaluated in full.

The transcript branch. Two mechanisms keep a question away from pixels, the transcript branch of §3.3 and the gate of §3.5, and Table 7 separates them. Whether the branch reaches a question at all is a property of the footage: it settles three quarters of Video-MME Long, half of HippoVlog and a third of LVBench, and nothing on the week-scale egocentric suites, whose transcripts run past its budget on every question. Under GPT-5 and measured against the reading loop on the same questions, the branch answers higher on Video-MME Long, 86.0 against 83.6, and on HippoVlog, 94.3 against 92.8, and lower on LVBench, 78.9 against 81.2. The remainder passes to the reading loop, where the gate fires on 9.3% of questions under GPT-5 and 22.5% under MiniMax-M3.

Transcript in context. The first row of Table 3 hands the answering function every transcript line before the question's query time in one call, with no memory and no retrieval. EgoLife's transcripts run past the 64k-token budget this row allows, so it thins them at a uniform stride that ignores the question, and the median question sees a quarter of its transcript.

Table 7: **The transcript branch.** Accepted is the share of questions the branch settles from the transcript before any loop runs, and Accuracy is how those questions score. The branch reaches no question on the egocentric suites, whose transcripts exceed its budget.

	GPT-5		MiniMax-M3	
Dataset	Accepted	Accuracy	Accepted	Accuracy
Video-MME (L)	76.1%	86.0	75.9%	78.5
HippoVlog	47.8%	94.3	50.9%	91.0
LVBench	31.9%	78.9	31.9%	70.4
EgoLife	0.0%	–	0.0%	–
Ego-R1	0.0%	–	0.0%	–
Overall	**38.9%**		**39.6%**	

C GATE ANALYSIS

This appendix reports how the gate's firing tracks the reader it monitors, measured with the transcript branch disabled so that every question passes through the reading loop and its gate.

Table 8 reports the result. Across the five benchmarks the gate's firing rate rises and falls with the reader's per-benchmark error rate, and the ratio between the two stays within 0.22 to 0.29, although the gate contains no per-benchmark parameter.

Overlap across models. In the deployed configuration the gate fires on 237 of the 4,194 questions under GPT-5 and on 569 under MiniMax-M3, and 77 questions fire under both. Two gates firing independently at those rates would share $237 \times 569/4{,}194 = 32.2$ questions, so the observed overlap is $2.39\times$ that expectation.

Discussion. The gate requires no calibration. A confidence-based trigger asks the model how certain it is and escalates below a fixed threshold, so its operating point rides on the scale the model puts on its own certainty, a scale documented to be miscalibrated (Kadavath et al., 2022); the ablation of §4.3 prices that dependence at 64.6 against 67.4 while firing on more questions. An equality test between two sampled answers carries no threshold to calibrate. The self-check of the transcript branch is a different instrument, judging whether a document read in full settles the question rather than scoring confidence in an answer, and the questions it commits answer at 86.0 against the configuration's 79.9 overall on Video-MME.

What the escalation gain contains. The reader scores 34.2 on the questions the gate forwards against 72.88 overall, and Table 9 follows those questions through the visual stages, to 53.6. How much of that climb needs new evidence? We draw the frozen answering function a third time over each question's own evidence history, spending one more call and receiving nothing new. This third draw reaches 43.0, so re-deriving recovers part of the climb, while Stage 2 is worth a further 8.1 points over it, 51 questions against 32 under McNemar, $p = 0.048$, and the deployed system 54 against 29, $p = 0.008$.

The gate fires only when two derivations disagree, so every escalated question carries two candidate answers. On 170 of the 237 the correct answer is one of the two, and on the remaining 67 it is neither. Those 67 are winnable only by returning a third answer, since nothing that re-picks between the two standing candidates can reach them. Table 10 counts how often each arm returns a third answer and how many of those departures fall on the 67. The third text draw departs from the pair on a schedule indistinguishable from chance. Stage 2 departs at twice the chance rate on exactly the questions both candidates missed, and the deployed system was correct on 20 of its 21 such departures. What the inspection contributes is therefore evidence that both standing hypotheses are wrong, rather than a further draw from the same distribution.

Where the gate spends the visual budget. Whether selecting those questions is worth doing depends on what the visual loop would have been worth on the questions the gate passed, so we hold

Table 8: **Gate adaptivity across datasets.** Firing tracks the reader's true error rate per dataset at a near-constant ratio, with no per-dataset parameters.

Dataset	Fire rate	Reader error rate	Ratio
EgoLife	10.0%	35.6%	0.28
Video-MME (L)	7.7%	26.3%	0.29
Ego-R1	8.3%	28.3%	0.29
HippoVlog	4.6%	20.4%	0.22
LVBench	7.7%	30.7%	0.25

Table 9: **The visual stages on the escalated subset.** Accuracy on the 237 questions the deployed GPT-5 configuration escalates, as each visual stage joins the reader's answer. The third text draw is a control: it spends one more call on the same evidence history and receives nothing new.

	Accuracy
Reader's answer a_1	34.2
Third text draw a_3, no new evidence	43.0
+ Stage 2	51.1
+ Stage 3	**53.6**

the gate open and run every EgoLife question through the loop, comparing against the deployed gate question by question.

Table 11 splits the questions by whether the deployed gate escalated them. The visual loop is worth 8.1 points on the questions the gate selects. On the questions it passes, forcing the loop through repairs as many answers as it breaks and is worth nothing. Holding the gate open spends $3.19\times$ the deployed gate's vision calls and reaches the same accuracy, 55.6 either way. The gate directs the visual budget at the questions that repay it.

Table 10: **Where a third answer lands.** Each arm is counted over the 237 escalated questions. Departures are the questions on which the arm returns an answer that is neither a_1 nor a_2, and the next column counts how many of those fall on the 67 questions whose correct answer is neither. Expected applies the 28.3% base rate to the departures, and p is a one-sided exact binomial test against that rate.

	Departures	Both wrong	Expected	p
Third text draw a_3	27	9	7.6	0.35
Stage 2	28	16	7.9	0.0013
Deployed system	36	21	10.2	0.00015

Table 11: **Where the visual loop pays.** EgoLife with MiniMax-M3 answering, comparing the deployed gate against an arm that holds the gate open so that every question runs the visual loop. Stage 1 replays from the deployed trace at zero model calls, so every change of answer is attributable to the visual loop. The gate escalates 29.8% of the questions in this configuration.

Questions	Reading alone	With the visual loop	Gain
The gate escalates	31.5	39.6	+8.1
The gate passes	62.4	62.4	+0.0

D QUERY-TIME COST

This section prices what READFIRST spends per question against a system that retrieves instead of reads. We reproduce WorldMM (Yeo et al., 2026) from its released code and its prebuilt EgoLife memory, answer with MiniMax-M3 on both sides, and meter every request at the API boundary. WorldMM answers 100 questions drawn from the 500 READFIRST answers. Both arms load memory built offline, and neither charges that construction to query time.

Table 12 reports what a deployment pays per question. READFIRST issues 14.9 calls against 19.0 and holds a memory 182 times smaller. WorldMM reads 83.9k tokens against 93.1k, the 10% that selecting three passages buys over reading a distilled transcript.

That token margin is measured in a unit that cannot see all of what WorldMM spends. Searching its visual memory encodes every query on a resident GPU encoder, and that compute returns no tokens for the meter to count, so the table prices a channel READFIRST does not run at all as though it were free.

We repeat the comparison on hour-scale video. Video-MME publishes no prebuilt memory, only its finest captions, so WorldMM's multiscale summaries, episodic and semantic triples and visual embeddings were built here with its own pipeline over 34 videos, and both systems then answered the same 102 questions.

Table 13 reports the result. READFIRST issues 5.0 calls against 24.4 and reads 17.7k tokens against 29.6k, and its memory takes 0.87 MB per video against 2.01 MB. The call ratio widens from EgoLife's $1.3\times$ to $4.9\times$ because 83% of these questions never leave the transcript: the cascade answers them in one reading pass and a review, while a retrieval agent runs its rounds whether or not the transcript already settled the question.

Table 12: **Query-time cost against WorldMM on EgoLife.** Both systems answer with MiniMax-M3, a call is one request the endpoint returned, tokens are the usage it reported, and the last row is what each system stores on disk for one subject. READFIRST issues fewer calls and holds a smaller memory, while WorldMM reads fewer tokens.

	READFIRST	WorldMM
Calls per question	**14.9**	19.0
Tokens per question	93.1k	**83.9k**
Query-time GPU	none	resident encoder
Memory on disk	**5.4 MB**	985 MB

Table 13: Query-time cost on Video-MME Long, MiniMax-M3 answering on both sides. Same conventions as Table 12: a call is one request the endpoint returned, tokens are the usage it reported, and the last row is what each system stores on disk for one video. READFIRST issues a fifth of the calls, reads three fifths of the tokens and holds a memory less than half the size.

	READFIRST	WorldMM
Calls per question	**5.0**	24.4
Tokens per question	**17.7k**	29.6k
Query-time GPU	none	resident encoder
Memory on disk	**0.87 MB**	2.01 MB

E BUILD-TIME COST

Both query-time tables hold memory construction outside the comparison, which is fair, since a deployment builds once and answers many times.

Table 14 prices building the memory for the 34 Video-MME videos. READFIRST spends 5,084 calls against 27,225, 226 per hour of footage against 1,209. What separates the totals is the graph: WorldMM extracts open information from every ten-second caption twice, once for entities and once for triples, then consolidates the triples a second time, while this system writes its scale summaries and stops. On EgoLife, whose released memory covers only the finest of four temporal granularities, extracting the other three for that subject cost a further 2,238 calls.

Table 14: **Build-time cost for the same 34 Video-MME videos, 22.5 hours of footage.** Both systems build with MiniMax-M3, and a number is the calls we metered at the endpoint, so a zero is a stage that costs nothing because the system never builds it. WorldMM's ten-second captions ship with its release and its visual embeddings are GPU work that issues no calls, so neither reaches our meter and its total is a floor. Neither system charges this construction to query time.

	READFIRST	WorldMM
Captions from frames	4,383	unmetered
Multiscale summaries and episodic triples	701	19,577
Semantic triples	0	7,648
Visual embeddings	0	unmetered
Metered calls	**5,084**	27,225
Metered calls per footage hour	**226**	1,209

F DISCUSSION

Two properties of the ablation behind Figure 1 govern how far its result carries.

Bounded visual access is a setting, not an assumption that sixteen frames exhaust a video. The experiment asks what happens when the visual evidence processed per question stays bounded while recording duration grows, the regime every long-video system occupies whenever query-time perception cannot scale with footage. We instantiate the regime at sixteen frames. The budget touches only part of the figure: the text-only curve uses no frames, and the collapse of frames alone is the arithmetic of §1. What it could distort is the margin of frames added to the language track. The crossover at which that margin turns negative may shift with a larger budget or a stronger selector, and our claim is not that sixteen frames universally suffice. We ask instead whether, under a fixed query-time visual budget, those frames should be attached to every question at all.

The comparison is paired. The text cell and the text-plus-vision cell receive the identical untruncated language track and differ only by the added visual observations. Sparse sampling can explain why frames repair fewer text errors as recordings lengthen. It does not by itself require those frames to overturn answers the text already had right. Both flows appear in the rates of Appendix A: repairs fall from 60.5% to 18.2% across the four tiers while reversals hold between 15.1% and 25.1%, and the two together carry the net marginal gain through zero.

Allocated per question instead, the visual loop is worth 8.1 points on the EgoLife questions the gate escalates and nothing on those it passes (Table 11), and perception run offline, distilled into the captions the reader retrieves, supplies the largest single gain of the component ablation (Table 3).

G LIMITATIONS

READFIRST requires one offline build, which transcribes the speech, distils captions from the frames, and consolidates both into a multi-scale textual memory. That cost is not particular to this method but a constraint the memory-based video systems share, and Appendix E prices it against the strongest baseline, which spends several times more. Perception is paid once and shared across questions, so the design suits settings where a recording is queried many times, and a single question against fresh footage does not amortise the build.