LEAP: LEARNED BLOCK-WISE EVIDENCE RETRIEVAL FOR LONG AUDIO-VIDEO PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hour-scale audio-visual question answering is constrained by a context dilemma: dense whole-recording encoding rapidly exhausts context limits, whereas uniform temporal compression severely dilutes fine-grained acoustic and visual evidence. We introduce LEAP, a framework where the model retrieves its own evidence without placing the whole recording in one context. LEAP divides a recording into fixed-duration blocks, applying a lightweight localization pass to each block to score short candidate windows. The highest-ranked windows are pooled and re-encoded in a single bounded answer pass. Consequently, the answer input and peak context remain independent of the recording duration. By decoupling evidence localization from reasoning, our framework can localize candidate temporal windows over pre-computed transcripts without decoding media frames, while preserving fine-grained visual and non-speech evidence by routing the final answering pass over raw audio-visual streams. LEAP trains both stages: a localization LoRA improves the selected windows, and an answer LoRA improves the answers read from the same windows. The block grid natively supports causal queries, enabling LEAP to support streaming inference without streaming-specific training. Across several AVQA benchmarks, LEAP improves over the Qwen3-Omni-30B-A3B baseline by 4.5–16.8%, and transfers to a second omni-modal backbone, MiniCPM-o 4.5, surpassing its published results by 3.1–13.0%.

1 INTRODUCTION

Long-form audio-visual question answering (AVQA) is challenging because the evidence is often scattered across a number of brief windows within a long recording. Processing the entire recording not only consumes substantial context and memory, but also fills the model's input with largely irrelevant content. Prior work addresses this along three axes: Selection methods keep a limited set of potentially relevant temporal regions before answering (Diao et al., 2025; Wang et al., 2026b; Shao et al., 2026). Compression methods keep the full recording but reduce its tokens or memory states (Tao et al., 2026a; Sun et al., 2026; Xin et al., 2026). Agentic methods revisit the source iteratively for each question (Wang et al., 2026b; Tao et al., 2025; Zhang et al., 2026b). However, selection made from a coarse view can drop the windows an answer needs, compression thins the detail of what it keeps, and agentic methods leave where to look to general-purpose models prompted at inference, none trained to localize evidence. We therefore ask: *Without full-context ingestion, can a model still pinpoint the few relevant minutes to answer a question?*

A fixed context makes coverage and density competing uses of the same tokens, which motivates visiting the timeline twice. Reading one timeline at two resolutions is an established remedy: a coarse pass decides where the evidence lies and a dense pass re-reads only those places. Prior systems scan the whole clip at low fidelity and zoom into the intervals they localize (Shen et al., 2025; Li et al., 2026), or descend recursively from long segments to short ones (Hannan et al., 2025). When the coarse pass reads the whole recording in one context (Li et al., 2026), its fidelity thins with duration and it stops at the model's position limit. The scan hands the dense pass a single interval (Li et al., 2026), so evidence scattered over several places is out of its reach.

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.

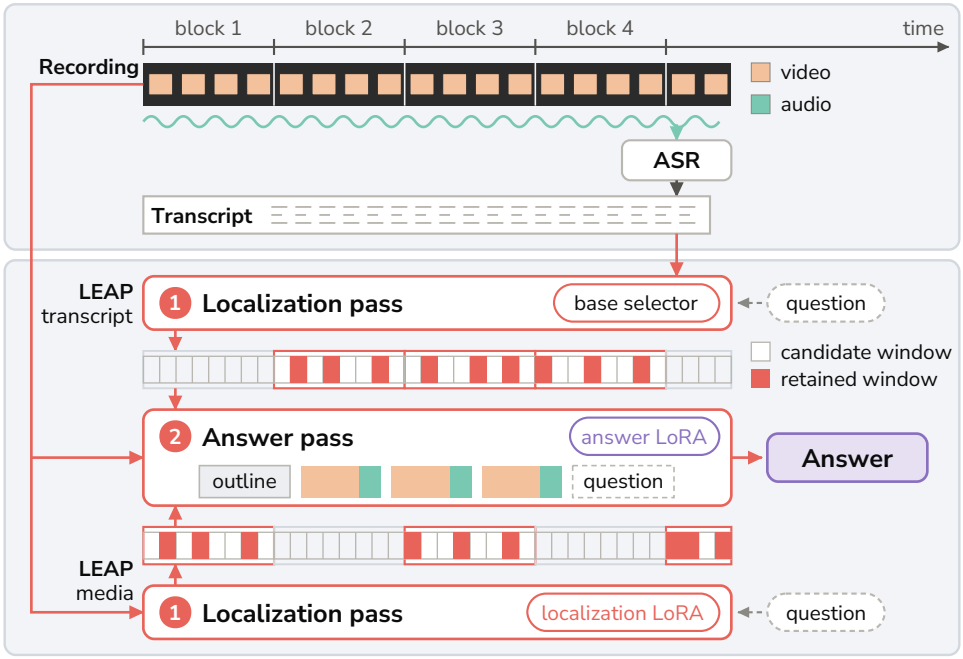

Figure 1: **Overview of LEAP.** *Top:* each recording is split into fixed-length blocks and transcribed once. *Bottom:* per question, a localization pass scores the candidate windows of every block (faint cells), from either the transcript with the base selector or the media with the localization LoRA; one answer pass re-reads the windows retained by block ranking (solid cells) from the recording.

We introduce LEAP, an evidence-retrieval framework for long-form AVQA where **the answering model retrieves evidence without placing the full recording in one context**, maintaining $\mathcal{O}(1)$ context and working memory. The temporal hierarchy is fixed by duration. The learning determines which windows are kept and how to answer from them. Specifically, the recording is partitioned into non-overlapping fixed-duration *blocks*, and each block is further partitioned into short *candidate windows*, which serve as the basic units of evidence selection. A lightweight, question-conditioned localization pass scores the windows within each block, and these scores are also used to rank and retain the most relevant blocks. LEAP selects only the highest-scoring windows and concatenates this small, bounded set into a single bounded *answer pass*. Each selected window is re-encoded at a fixed per-window context, dense enough to preserve the local details needed for answering.

Across several AVQA benchmarks, LEAP improves accuracy over the Qwen3-Omni-30B-A3B baseline by 4.5–16.8%. Localization training improves the windows the retrieval stage selects, and answer training improves the answer read from those windows. LEAP also transfers to MiniCPM-o 4.5 Cui et al. (2026), surpassing its published results by 3.1–13.0%.

We summarize our contributions as follows.

- **Evidence retrieval without a whole-recording read.** We introduce LEAP, which narrows the full-length recordings to a bounded set of evidence-rich windows at per-pass context and working memory $\mathcal{O}(1)$ in duration.
- **The transcript as a second scanning channel.** Decoupling evidence localization from reasoning lets the same block grid be searched on pre-computed transcripts without decoding media frames; only the windows it selects are re-read as audio and video, so the two channels share the block grid and the answer pass and differ in what is scanned. Routing the answer pass over the raw audio-visual stream preserves the fine-grained visual and non-speech evidence a transcript misses.
- **Historical retrieval as the stream arrives.** The block grid natively supports causal queries, enabling LEAP to support streaming inference without streaming-specific training (§4.6).

Localization pass selects windows over the media blocks, or over an online transcript index, up to query time, and the answer pass re-reads them from raw media.

2 Related Work

Long-form audio-video question answering. Omni-modal models such as Qwen3-Omni (Xu et al., 2025) jointly reason over text, video and audio. However, their finite context windows make hour-scale reasoning difficult. Existing approaches address this problem in three main ways. *Selection-based methods* retain frames, clips, or intervals before reasoning (Diao et al., 2025; Shao et al., 2026; Pan et al., 2025; Shen et al., 2025; Zhang et al., 2026a; Hannan et al., 2025; Li et al., 2026). *Compression-based methods* preserve coverage while reducing visual tokens, representations, or memory states (Gong et al., 2025; Tao et al., 2026a; Ding et al., 2026; Xin et al., 2026; Sun et al., 2026; Li et al., 2024; Shu et al., 2025). *Agentic methods* revisit the source iteratively according to the question (Tao et al., 2025; Xing et al., 2026; Wang et al., 2026b; Zhang et al., 2026b; Zhu et al., 2026). Streaming understanding adds a causal constraint: only the recording prefix is readable when a question is asked. Streaming systems are built to ingest each frame once and answer from what they keep, either a KV cache of the stream (Di et al., 2025; Chen et al., 2026), a fixed-size memory (Zhang et al., 2025; Zeng et al., 2026), or a textual memory of distant history (Jiang et al., 2026). ShallowStream instead uses a shallow-layer KV cache as an index and re-processes the frames it ranks through all layers (Hao et al., 2026). These three families share one tension: within a fixed context, covering the recording and resolving fine evidence compete for the same tokens. Selection can therefore drop the windows an answer needs, compression thins the detail in the windows it keeps, and agentic revisiting eases the tension by re-running its media work for every question.

Text as a retrieval channel. Text provides an efficient way to search long recordings without repeatedly processing the original media. Caption-based pipelines aggregate descriptions of short clips to answer long-range video questions (Zhang et al., 2024), while document-retrieval systems convert video into searchable text that is ultimately consumed by the answering model (Ma et al., 2025). Other systems search a text index built over the recording while keeping the media indexed beside it, so a query that starts in captions or transcripts can drop back to frames (Yin et al., 2026; Zhang et al., 2026b; Wei et al., 2026). Others provide retrieved ASR and OCR text to the model alongside the video (Luo et al., 2026). Transcripts also serve as retrieved evidence for an omni-modal agent (Zhu et al., 2026), and a long-audio planner searches timestamped streams derived from the audio, transcript among them (Someki et al., 2026). Subtitle-based keyframe selection performs comparably to visual-search (He et al., 2026b), and query-conditioned gating can select the retrieval channel or depth for each question (Wang et al., 2026a; Xue et al., 2025). The caption-based, document-retrieval and long-audio planning pipelines hand their answering model text alone, never the recording itself (Zhang et al., 2024; Ma et al., 2025; Someki et al., 2026).

3 Methodology

3.1 Problem Formulation and Overview

Given a temporally aligned video stream $\mathcal{V}$, audio stream $\mathcal{A}$, a question q and its answer options O, LEAP generate an answer y without placing the complete hour-scale recording in the model's context window. As shown in Fig. 1, LEAP has two stages: one *localization pass* per fixed-duration audio-video block, scoring that block's candidate windows (§3.2.1), and one bounded *answer pass* re-encoding the high score windows (§3.2.2). Algorithm 1 lists the full inference procedure.

The *block grid* is fixed by the clock. The stream is partitioned into non-overlapping *blocks* $x_1, \ldots, x_M$ of duration $\Delta = 600$ s, $M = \lceil T/\Delta \rceil$ of them for a recording of length T. Each block is tiled into non-overlapping *candidate windows* of duration $\delta = 75$ s, so a full block carries $K = \lceil \Delta/\delta \rceil = 8$ of them and the final block, if shorter, only its $K_m \leq K$ windows that hold content. The ranking retains $B = 3$ blocks, and inside each retained block the shortlist keeps $W = 3$ windows, so at most $BW = 9$ windows enter the answer pass. Appendices C.1 and C.4 ablate the block count, the window count and the window width. The two modalities are treated asymmetrically in the localization pass: video is compressed to a small set of visual tokens by sparse frame sampling, while all 600 s of audio

are encoded at the backbone's native rate (Appendix A.3). The audio keeps its full temporal support, and the visual sequence stays small enough for one inexpensive pass per block.

3.2 Framework

LEAP runs on two omni-modal models, Qwen3-Omni-30B-A3B (Xu et al., 2025) and MiniCPM-o 4.5 (Cui et al., 2026), whose weights θ_0 stay frozen. An audio encoder turns each second of the waveform into 13 tokens. A vision encoder turns the sampled frames into visual tokens. The language model reads the two token streams interleaved with the text of the prompt and produces text. A block x_m encoded this way at the localization-pass media rate is the z_m the localization pass reads. The trained parameters are the two LoRA adapters ϕ_{loc} and ϕ_{ans}, each a rank-16 update to the query, key, value and output projections of every self-attention layer of the language model. This section's configuration values and training objective are Qwen3-Omni's; Table 5 lists MiniCPM-o's.

3.2.1 Stage I: The Localization Pass and Window Scoring

The frozen backbone with the localization LoRA adapter (Hu et al., 2022) processes each block exactly once. The block's K_m candidate windows are listed as lettered options in a fixed prompt; the localization pass reads the next-token logit of each option letter, and both scores use these logits,

$$(\ell_{m,1}, \ldots, \ell_{m,K_m}) = \mathcal{F}_{\theta_0,\phi_{\text{loc}}}(z_m, q)\,, \qquad r_{m,k} = \sigma(\ell_{m,k})\,, \qquad g_m = \max_{k\in\{1,\ldots,K_m\}} r_{m,k}, \quad (1)$$

where z_m is the compressed representation of block x_m and $\sigma(\cdot)$ the logistic sigmoid. A block is ranked by a monotone rescaling of its strongest candidate-window logit, a *ranking* score read on the logit scale the passes share. The aggregation is a max (§3.3), ablated against a mean in Appendix C.6. The localization pass emits option-letter logits only, with no decoded answer and no cross-block key–value cache, so the peak context and memory of one localization pass depends on a single block.

3.2.2 Stage II: Block Ranking and Answer Pass

After all blocks have been scanned, we retain the $\min(B, M)$ highest-scoring blocks under g_m, restored to chronological order before answer generation. The whole pipeline costs $N_{\text{fwd}} = M + 1$ passes per question, with no dependence on B (Appendix A.4).

Within each retained block the shortlist keeps its $\min(W, K_m)$ highest-scoring windows under $r_{m,k}$. Every block's shortlist is formed during its own localization pass, from the window scores Eq. 1 has already produced. The retained blocks' shortlists are pooled into the evidence of the answer pass.

Only these selected windows are reloaded from the original recording. Sorted by absolute time into windows $(J_1, \ldots, J_R)$ with $R \le BW$, they form the bounded answer-pass sequence

$$\big[\omega;\, c_0;\, \tau_1;\, \text{Enc}(J_1);\, \cdots\, ;\, \tau_R;\, \text{Enc}(J_R);\, q;\, O\big], \quad (2)$$

where ω is the transcript outline introduced below, c_0 is a fixed instruction stating that the segments are discontinuous and must be reasoned over jointly, τ_i carries window i's absolute timestamp, and $\text{Enc}(J_i)$ encodes the synchronized video and audio of the ith window's temporal support J_i. Each retained window is re-read from the source at the *per-window context*: 32 video frames per 75 s window plus that window's audio at the backbone's native rate.

The transcript outline. The answer pass also reads ω, a text outline of the whole recording, built from a single transcription made once, before any question is asked: one line per minute of speech, each carrying its minute mark, capped at 4,000 tokens. A transcription longer than the cap is filled question-first, the minutes matching the question and its options admitted ahead of the rest (Appendix A.3). A recording without speech contributes no outline.

The backbone swaps the localization adapter ϕ_{loc} for the answer adapter ϕ_{ans}; each is trained from the frozen base backbone on its own data, and the two are never stacked. The answer pass generates over all selected windows jointly, so evidence from different blocks interacts in one bounded context.

3.3 Localization Adapter

The localization adapter is trained on this task, on questions derived from LongVALE (Geng et al., 2025), whose event annotations supply the evidence span. Every training clip fits in one block. For

a clip with annotated evidence span $J^\star$, the supervised letter is the candidate that best covers it. Supervision is the cross-entropy over the same option-letter logits the localization pass reads,

$$\mathcal{L}_{\text{loc}} = -\sum_{(z,q,J^\star)\in\mathcal{D}_{\text{loc}}} \log \frac{\exp(\ell_{k^\star})}{\sum_{k=1}^{n}\exp(\ell_k)}, \qquad k^\star = \arg\max_k |J_k \cap J^\star| \,. \tag{3}$$

Here z is a training clip encoded at the localization-pass media rate, q its localization question, and $J^\star$ its annotated evidence span. The sum runs over the n lettered candidates $J_1, \ldots, J_n$ of one supervision level (Table 4). These candidates are set by the training clip rather than by δ: every training clip is shorter than one block, and tiling it at δ would leave most clips with fewer than the eight letters a deployed localization pass reads. A longer clip is therefore first cut into eight equal candidate windows, carrying whole-clip audio and sparse video as a deployed localization pass does, so its first level reads the same eight letters; a second, training-only level covers the finer windows inside the one that holds the evidence. A shorter clip is tiled directly into at most eight finer windows.

Training applies the cross-entropy objective of Eq. 3 to the single target $k^\star = \arg\max_k |J_k \cap J^\star|$, the window that overlaps the evidence most. MiniCPM-o 4.5's selector is instead trained with overlap-fraction BCE, a binary cross-entropy that grades every window by its overlap with the evidence (Table 5); on Qwen3-Omni that objective ranks blocks worse across passes, yet does not yield a statistically detectable difference in final accuracy (Appendix C.6). Inference scores a block by its maximal candidate-window posterior, $g_m = \max_k \sigma(\ell_k)$. Replacing the max by a mean discards the margin by which the block's best window stands above the other windows of its pass, the component the block ranking runs on, and costs most of the evidence coverage (Appendix C.6).

3.4 Answer Adapter

The answer adapter is trained via cross-entropy over the answer token and the end-of-turn token, conditioned on an evidence sequence $\mathcal{X}$ of the same form but without the outline ω,

$$\mathcal{L}_{\text{ans}} = -\frac{1}{|y|}\sum_{t=1}^{|y|} \log p(y_t \mid y_{<t}, \mathcal{X}) \,. \tag{4}$$

The sequence of Eq. 2 is the inference-time instance of $\mathcal{X}$, with ω added. Each training instance stitches the gold evidence segment with distractor segments drawn from *other* recordings, rendered in the same form as the retrieved sequence: timestamped raw audio-video segments, one of which carries the evidence. Full setup is in Appendix A.2.

3.5 Advantages

Bounded context and peak memory. Either pass reads a context of fixed size: the localization pass reads one block, and the answer pass at most BW selected windows together with the outline and the question, so peak context and memory per pass are $\mathcal{O}(1)$ in T. One pass over the whole recording instead costs ≈47k tokens per hour: the backbone's 65,536-token position limit is exhausted at roughly 81 minutes. Appendix A.4 details the token accounting.

Localization training within one block. The objective of Eq. 3 is defined within one block, so training, like inference, never reads more than one block, whereas selectors trained by reinforcement from the answer (Pan et al., 2025; Li et al., 2026) carry the whole recording and a live answering model in every update. Its supervision marks where the evidence lies, never answer correctness, so the selector never learns which inputs one answering model gets right (Appendix A.5).

Coverage and detail on separate passes. LEAP separates covering the recording from resolving its evidence: the localization pass covers every block, and the answer pass spends its context on the few retained windows, at a per-window context that does not change with T. Whichever channel selects those windows, media or transcript, the answer pass re-encodes the same raw audio-video windows and reads the transcript only as a short outline. What a transcription misses, unspoken visual detail or non-speech audio, therefore stays readable at answering time.

Table 1: **Main results on Qwen3-Omni-30B.** Accuracy (%). Upper block: closed-source systems (°) as published. Middle block: open systems as published, then our backbone; * = the published number for our backbone; *caption cascade* = audio and visual captions read by a text-only model; ‡ = the benchmark's official input recipe re-run on our stack, or an official-style whole-clip run where none is runnable. Lower block: *LEAP without retrieval* = LEAP's answer LoRA reading the whole clip in one pass (on VideoOdyssey, its audio-montage recipe); *LEAP without answer training* = LEAP's localization LoRA also answering the selected windows; — = no number. TraceAV = the mean over its twelve general sub-tasks. **Bold** = best, underline = second best in each column.

System	TraceAV	LVOmni	VideoOdyssey (1–4h)	MMOU
gemini-3-flash-preview°	**62.3**	**59.0**	44.3	—
Qwen3.5-Omni-Plus°	—	—	43.0	—
Ming-Flash-Omni-2.0	51.7	34.6	—	—
Caption cascade into Qwen3-235B	—	—	—	47.9
Qwen3-Omni-30B, as published*	48.4	35.8	28.7	54.1
Qwen3-Omni-30B, with ASR transcript*	—	42.2	—	—
Qwen3-Omni-30B, official recipe	56.4‡	40.7‡	36.9‡	56.5‡
Whole clip, uniform, frame-matched	53.6	40.4	—	57.8
LEAP without retrieval	54.8	42.9	41.9	61.8
LEAP without answer training	60.1	43.0	46.7	61.4
LEAP	60.9	45.8	**53.7**	**66.3**

4 Experiments

4.1 Settings

We evaluate LEAP on four main benchmarks: TraceAV-Bench (Feng et al., 2026), LVOmniBench (Tao et al., 2026b), VideoOdyssey-AV (He et al., 2026a) (videos all longer than 60 minutes), and MMOU (Goel et al., 2026). We also evaluate OmniVideoBench (Li et al., 2025), included in the paired ablations of this section, and on three video-only benchmarks, LVBench (Wang et al., 2025), CG-Bench mini (Chen et al., 2025) and Video-MME (Fu et al., 2025), reported in Appendix E.1. The block configuration is unchanged across all datasets and carried as is to StreamArena (Zhang et al., 2026c) in §4.6. On the four main benchmarks, CG-Bench and StreamArena the answer pass over the selected windows carries the transcript outline of §4.5; the selector comparisons of Figure 3 and Table 11 answer without it. The *base selector* is the same selection run by the *base model*, the backbone with no adapter mounted. An effect is *significant* when its paired video-clustered interval excludes zero (Appendix B.1). Appendix B.6 reports the contamination audit.

4.2 Main Results

Table 1 compares LEAP against same-stack baselines using either the official input recipe or an official-style whole-clip run (Appendix B.3). Our method leads significantly on all four. LEAP also leads the whole clip read, sampled uniformly and frame-matched baselines. Figure 2(a) extends the comparison beyond the four main benchmarks, reading LVBench and Video-MME with audio removed as their official protocols do, and our method leads on all of them.

4.3 Retrieval Stage

Figure 3 shows localization training raises *evidence coverage*, the share of questions whose annotated evidence the selection retains: substantially in the answer-pass windows, and in the retained blocks most on VideoOdyssey. How much of that added coverage turns into accuracy differs across benchmarks: LEAP has almost no evidence coverage left to gain on TraceAV, and by far the most on VideoOdyssey. The trained selector also beats an untrained ranking score: an off-the-shelf retriever or the confidence of an answer read on each block put in its place, everything else held, covers less of the evidence and answers less accurately, pooled over the benchmarks of Table 11.

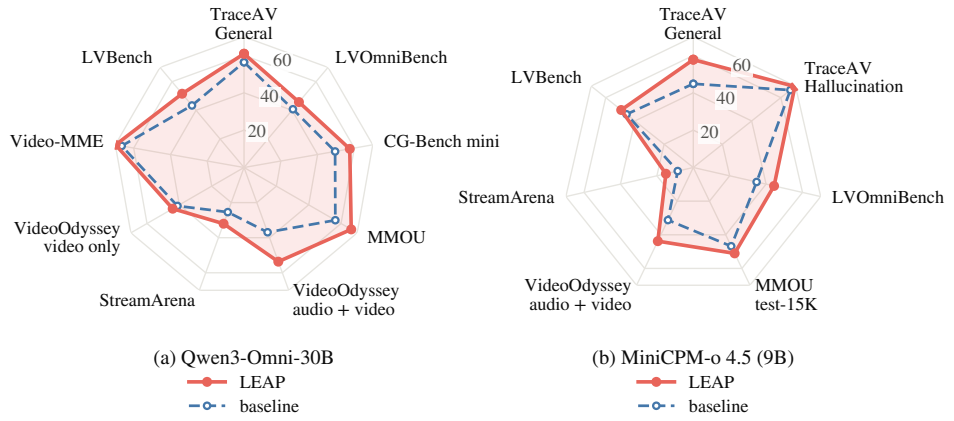

Figure 2: **Across benchmarks and backbones.** Solid = LEAP, dashed = baseline: *(a)* the same backbone on the whole recording (the causal prefix on StreamArena); *(b)* the backbone as published (official-style on our stack for VideoOdyssey and LVBench, native streaming on StreamArena).

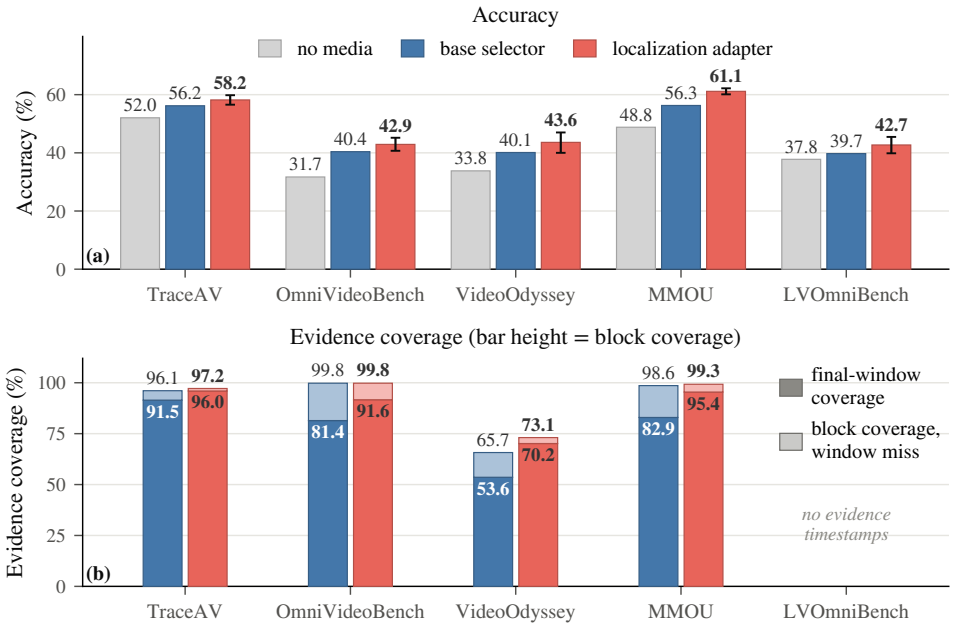

Figure 3: **The retrieval stage ablated.** *no media*: stem and options only. *(a)* Accuracy; the localization adapter's whisker (paired, vs. base selector) clears the base-selector bar where significant. *(b)* Evidence coverage: full bar = in a retained block, saturated = in an answer-pass window; OmniVideoBench on its timestamped subset, MMOU where both selectors stored a selection.

LEAP gains from where it places its windows: it answers significantly more accurately on TraceAV and VideoOdyssey than when the same windows are redrawn at random positions over the recording (Appendix C.5). It also gains from its block ranking: on VideoOdyssey it is significantly ahead of equally spaced blocks in both accuracy (Figure 7) and evidence coverage (Appendix C.6).

4.4 Answer Stage

Table 8 fixes the answer LoRA and changes only what it reads: the whole clip, or LEAP's windows. LEAP leads on all four benchmarks. Part of that lead comes from clips too long to encode whole. Without those questions, LEAP keeps most of its lead on LVOmniBench, still significant. Against the frame-matched whole clip, which reads as many frames as LEAP but is answered by the base model,

Table 2: **Transcript-based retrieval against the media channel.** Accuracy (%). Both channels feed the deployed answer pass. *Localization tokens* count the localization passes only; *seconds* = question-to-answer wall clock on one idle GPU over a duration-stratified sample, transcription charged to neither. TraceAV = accuracy over all its questions, hallucination sub-tasks included.

System	TraceAV	LVOmni	VideoOdyssey (1–4h)	MMOU
Media channel	63.6	45.8	53.7	66.3
Transcript channel	64.6	46.5	52.3	66.2
Localization tokens per question, media channel ÷ transcript channel				
Media ÷ transcript	5.41×	4.74×	6.21×	4.93×
Query-time seconds per question (median)				
Media channel	12.4	15.6	30.0	5.5
Transcript channel	7.6	11.0	18.6	3.4

LEAP leads significantly on TraceAV, LVOmniBench and MMOU (Appendix B.4). LEAP transfers to MiniCPM-o 4.5 with the same block grid and backbone-specific adapter training (Table 5). LEAP raises evidence coverage and accuracy significantly on VideoOdyssey and MMOU (Appendix D).

4.5 Audio and Transcripts

LEAP reads audio in both passes. On TraceAV, silencing only the *localization pass* changes the selected windows substantially. Muting only the answer pass costs accuracy significantly on the hearing-required and cross-modal question classes (Table 16b).

Transcript channel. The media channel spends one localization pass on the video and audio of every block, speech often provides sufficient discriminative information. We therefore transcribe each recording once, before any question, and build a second channel that selects windows from the transcript alone. Each candidate window is represented by the transcript of its span and scored by the base selector, with the media channel's prompt and the same block-then-window selection.

Both channels hand their windows to the deployed answer pass (Table 2). The transcript channel spends a fraction of the media channel's localization tokens and answers faster on every benchmark. The more of the annotated evidence is spoken, the more of it the transcript channel retains relative to the media channel (Appendix C.2). LEAP deploys the media channel throughout; where a recording's evidence is spoken, the transcript channel is the cheaper substitute.

Transcript outline. LEAP's answer pass also reads a minute-stamped outline of the same transcript. Table 13 replaces the deployed outline by one that thins every minute equally, or removes it. Removing the outline costs LEAP accuracy significantly on TraceAV, VideoOdyssey and MMOU. On the questions whose outline is cut, LEAP's question-first outline beats one thinning every minute equally, significantly on VideoOdyssey. LEAP's gain from the outline is largest per question where the retained windows miss the evidence (Appendix C.3).

4.6 Duration and Evidence Distance

LEAP's accuracy shows no systematic collapse as videos get longer (Figure 4), and on MMOU it rises with duration. LEAP and a single whole-recording pass spend their context differently as the duration T grows. A baseline that fits the whole recording into one fixed context must spread that context over all of it, so what it reads per minute of source falls as $1/T$. LEAP reads every retained window at the same per-window context at any duration. A strong baseline is the context-filled montage (Appendix B.1), which packs the whole recording into one bounded pass. LEAP leads significantly in every quintile, including the one where the montage keeps almost all the audio (Figure 15).

Causal access. StreamArena (Zhang et al., 2026c) asks each question at a moment of a stream, lets the model read only what came before that moment, and bins the questions by how far back the evidence lies. LEAP's block grid runs on it unchanged, with no streaming training. The answer pass reads the retained windows with the base weights and the transcript outline, and always includes the

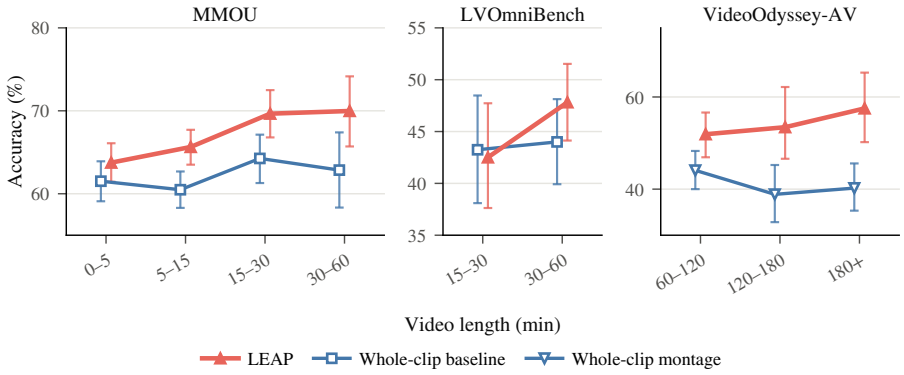

Figure 4: **Accuracy vs. video length.** All panels share one grid of length buckets, and each shows the buckets it populates. Error bars are each line's video-clustered 95% interval.

Table 3: **Historical retrospection under causal access on StreamArena, Qwen3-Omni-30B.** Accuracy (%). Bins = minutes from the query time back to the evidence. *Whole prefix* reads everything before the query time in one pass; *Uniform windows* fills the same context with windows spread equally over it; *LEAP – media* and *LEAP – transcript* select the windows from the media blocks or from the transcript index built as the stream arrives, and answer them from the raw media. **Bold** = best, underline = second best in each column.

		evidence distance (min)			
System	HR avg	≤5	5–15	15–30	>30
Whole prefix	25.4	22.1	34.2	25.5	12.6
Uniform windows	20.2	23.2	25.8	17.9	9.3
LEAP – media	**32.0**	**26.5**	**41.2**	29.3	23.6
LEAP – transcript	30.1	24.3	36.1	**30.4**	**24.7**

window ending at the query time (Table 5). The block grid is fixed by the clock and every candidate window is scored inside its own block (Appendix B.5). LEAP significantly leads the same backbone reading the whole prefix in one pass and windows spread equally over the prefix (Table 3).

The transcript channel of §4.5 suits a live stream. The stream is transcribed in time order with no lookahead, a question selects windows from the transcript available at its query time, and the answer pass re-reads those windows from the raw media. Its windows contain the annotated evidence about as often as those of the media channel. Answering them from their transcript instead of the raw media costs accuracy significantly in all but the farthest distance bin. On the 9B MiniCPM-o 4.5, LEAP also beats the model's native streaming mode significantly, overall and in the second distance bin.

5 Conclusion

Long-form audio-visual question answering is limited by a fixed context: spread over hours, it leaves too little detail per minute. We introduced LEAP, an evidence-retrieval framework that reads the recording in bounded pieces: every pass reads one fixed-duration block or a bounded set of retained windows, so peak context and memory are independent of duration. Across four benchmarks LEAP improves over each benchmark's official-style recipe on the same backbone by 4.5–16.8%, shows no systematic collapse with duration, and transfers to a second omni-modal backbone. The same block grid searches a timestamped transcript with 4.7–6.2× fewer localization tokens than the media channel, and applies to a causal prefix without streaming training. Open directions include stronger retrieval scores within the bounded-cost structure, and an answer pass that can tell when its retained windows miss the evidence and return to the scan for more.

AI USE STATEMENT

Among the tasks with required disclosure, we used generative AI tools to propose and refine hypotheses, give feedback on research methodology and experiments, implement methods, clean datasets, support qualitative data analysis, interpret results and assist with translation. We have not used them to generate synthetic data sets or to develop theoretical models or conceptual frameworks, and formulating mathematical claims and writing proofs are not applicable to this work. Additionally, we used them for the tasks with recommended disclosure below.

Writing. We used them to draft and polish prose, captions, figures and LaTeX; the authors edited and approved every sentence.

Retrieval and discovery. We used them to search for and summarize related work; an author read and verified every cited work in the original.

Research ideation and execution. We used them for design discussions, as coding assistants, and for labeling in Appendices A.3 and B.5. No reported result was written or altered by a model.

We have reviewed all AI-assisted work, and we take responsibility for the final content of this paper, including its text, claims and artifacts.

REPRODUCIBILITY STATEMENT

LEAP is specified in §3 and step by step in Algorithm 1. Both adapters, their training pools and the inference settings are listed in Table 4, and those of the second backbone in Table 5; Appendix A.3 gives how the localization pool is built, the optimizer, the prompts, the letter read-out and the transcript rendering. Appendix B.1 fixes the evaluation protocol of our runs, including the denominators and the clustered confidence intervals; Appendices B.3 and B.5 describe how the same-stack baselines and the causal-access protocol are run, and Appendix B.6 audits the training sources against the benchmarks. All benchmarks and training corpora are public releases. The supplementary material contains anonymized source code for data building, training, both inference passes, transcription and scoring, with the training configurations, a script that runs the full pipeline with the settings of the paper, and worked examples that carry benchmark questions through both passes.

REFERENCES

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. Omnivideo-100k: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, and Ming Lu. Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3120–3128, 2026.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, TaoZhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video KV-cache retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8g9fs6mdEG.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. Temporal working memory: Query-guided segment refinement for

enhanced multimodal understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3393–3409, 2025.

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Hengyi Feng, Hao Liang, Mingrui Chen, Bohan Zeng, Meiyi Qiang, Zhengyang Zhao, Zimo Meng, Zeang Sheng, and Wentao Zhang. Traceav-bench: Benchmarking multi-hop trajectory reasoning over long audio-visual videos. *arXiv preprint arXiv:2605.07593*, 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, 2025.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*, 2026.

Chao Gong, Depeng Wang, Zhipeng Wei, Ya Guo, Huijia Zhu, and Jingjing Chen. Echoingpixels: Cross-modal adaptive token reduction for efficient audio-visual llms. *arXiv preprint arXiv:2512.10324*, 2025.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19012–19022. IEEE, 2025.

Jitai Hao, Ke Yang, Qiang Huang, and Jun Yu. Shallowstream: Index shallow then answer deep for streaming video understanding, 2026. URL https://arxiv.org/abs/2609.02780.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514. IEEE, 2024.

Haichen He, Jiayi Zhou, Sifeng Shang, Yihan Hu, Yuanhan Zhang, and Kaiyang Zhou. Videoodyssey: A benchmark for ultra-long-context and omni-modal video understanding. *arXiv preprint arXiv:2605.22907*, 2026a.

Jianxiang He, Meisheng Hong, Jungang Li, Weiyu Guo, Xuming Hu, and Hui Xiong. Vsi: Visual-subtitle integration for keyframe selection to enhance long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9003–9012, 2026b.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xinru Jiang, Lin Zhao, Xi Xiao, Yunbei Zhang, Janet Wang, Chenrui Ma, Haolin Li, Yanzhi Wang, Yifan Gong, and Octavia Camps. Dynamic hub-and-spoke memory for streaming video understanding. *arXiv preprint arXiv:2608.30294*, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

Yixuan Li, Guangzhi Sun, Yudong Yang, and Chao Zhang. video-salmonn-r^3: Learning to rewatch, reask, and reanswer for efficient video understanding, 2026. URL https://arxiv.org/abs/2606.24477.

Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, et al. Cachegen: Kv cache compression and streaming for fast large language model serving. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 38–56, 2024.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946. IEEE, 2025.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *arXiv preprint arXiv:2511.05489*, 2025.

Mingchen Shao, Hang Su, Wenjie Tian, Bingshen Mu, Zhennan Lin, Lichun Fan, Zhenbo Luo, Jian Luan, and Lei Xie. Listening with time: Precise temporal awareness for long-form audio understanding. *arXiv preprint arXiv:2604.22245*, 2026.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in. *arXiv preprint arXiv:2512.14273*, 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169. IEEE, 2025.

Masao Someki, Chien-yu Huang, Siddhant Arora, Samuele Cornell, Markus Müller, Nathan Susanj, Rupak Vignesh Swaminathan, Grant Strimel, Jing Liu, and Shinji Watanabe. Planrag-audio: Planning and retrieval augmented generation for long-form audio understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26167–26183, 2026.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 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.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omniagent: Audio-guided active perception agent for omnimodal audio-video understanding. *arXiv preprint arXiv:2512.23646*, 2025.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17682–17692, 2026a.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, et al. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms. *arXiv preprint arXiv:2603.19217*, 2026b.

Shaoguang Wang, Weiyu Guo, Ziyang Chen, Xuming Hu, and Hui Xiong. Where to focus: Query-modulated multimodal keyframe selection for long video understanding. *arXiv preprint arXiv:2604.17422*, 2026a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026b.

Fan Wei, Siru Zhong, Runmin Dong, Miao Yang, Zhaoyang Luo, and Haohuan Fu. Ground, cover, and refine: Evidence-centric frame selection for long-video question answering, 2026. URL https://arxiv.org/abs/2608.01660.

Zijie Xin, Jie Yang, Ruixiang Zhao, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. Stage-adaptive token selection for efficient omni-modal llms. *arXiv preprint arXiv:2605.20035*, 2026.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, and Pheng-Ann Heng. Native active perception as reasoning for omni-modal understanding, 2026. URL https://arxiv.org/abs/2606.19341.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 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. *arXiv preprint arXiv:2506.13589*, 2025.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. *arXiv preprint arXiv:2601.13719*, 2026.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pp. 84154–84179, 2025.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069. IEEE, 2025.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026a.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026b.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding. *arXiv preprint arXiv:2608.05703*, 2026c.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

Yifan Zhu, Xinyu Mu, Tao Feng, Zhonghong Ou, Yuning Gong, and Haoran Luo. Omnirag-agent: Agentic omnimodal reasoning for low-resource long audio-video question answering. *arXiv preprint arXiv:2602.03707*, 2026.

A The Method in Detail

A.1 The inference algorithm

See Algorithm 1.

Algorithm 1 LEAP inference.

Require: video V with its audio (any duration), question q, options O

$\omega \leftarrow \textsc{Outline}(V, q, O)$ ▷ one transcription per recording, its minutes ranked for the question
$x_1, \ldots, x_M \leftarrow \textsc{TileNonOverlapping}(V, 600\text{s})$
for $m = 1, \ldots, M$ **do** ▷ exactly one localization pass per block
  $\ell \leftarrow \textsc{LocalizationPass}(x_m, q)$ ▷ up to 8 window letters
  $\mathcal{U}_m \leftarrow \textsc{Top}(\ell, W{=}3)$; $g_m \leftarrow \max_{r \in \mathcal{U}_m} \sigma(\ell_r)$
end for
$\mathcal{S} \leftarrow \textsc{Top}(\{g_m\}, B{=}3)$ ▷ selected blocks
$\mathcal{U} \leftarrow \textsc{TimeSortedUnion}(\{\mathcal{U}_m : m \in \mathcal{S}\})$ ▷ $\leq BW{=}9$ windows by construction
return $\textsc{Answer}(\textsc{PackWindows}(\mathcal{U}, \omega, q, O))$ ▷ per-window context, one bounded pass

A.2 Training and inference configuration

Table 4 lists both adapters' configuration and the deployed inference settings; Table 5 lists where MiniCPM-o 4.5 and the causal-access protocol of Appendix B.5 depart from them. On MiniCPM-o 4.5 the answer pass reads fewer frames per second than the localization pass, ten per window being where this backbone's accuracy stops rising on long recordings. Re-answered with the localization adapter mounted over the same retained windows, more frames per window, up to one per second, do not raise accuracy on LVOmniBench or OmniVideoBench, and one per second lowers it.

A.3 Data, prompts, and renderers

Localization corpus. Every LongVALE event segment yields one question: the stem asks in which time window the segment's caption occurs, the segment's span is the annotated evidence span $J^\star$, and the clip is its source video, which never exceeds one block. The candidate windows are enumerated at training time from the evidence span and the clip duration on the grids of Table 4, and the target at each level is the candidate overlapping the evidence span most, the earlier one on ties. An evidence span that starts inside the clip's first 3 s or first 5% is dropped. The quality filter is the base model: shown one candidate window at a time, it is asked whether the evidence for the event appears in that segment, and a question is kept only when it says yes on the target candidate and no on every other candidate; an unparseable reply counts as no. The corpus keeps 40,284 of 69,630 questions, over 4,735 source videos. Videos are split first, by a hash of their identifier, one in ten to validation.

Optimization. Both adapters are trained by one trainer: AdamW with $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$ and decoupled weight decay 0.01, learning rate 10^{-4} under a cosine schedule without warmup, gradient norm clipped at 1, bf16 with gradient checkpointing, two data-parallel ranks, the gradient averaged over the items of a step. The localization adapter's two supervised levels each contribute one

Table 4: **Configuration of the two adapters.**

Backbone and adapters	
Backbone	Qwen3-Omni-30B-A3B-Instruct, frozen in both stages
LoRA rank / scaling	$r = 16$ / $\alpha = 32$, identical for both adapters
Attachment points	attention query, key, value and output projections
Composition	sequential, one adapter resident at a time, never stacked
Training hardware	H100, bf16 throughout
Localization adapter ϕ_{loc}	
Objective	cross-entropy over the option-letter logits against the best-covering candidate, Eq. 3, at each supervision level listed below
Data	LongVALE-derived localization questions, filtered by the base model (Appendix A.3)
Training clip length	30–598 s; one 600 s block each
Supervision grid	clips up to 240 s: one level of 30 s windows; longer clips: two levels, first 8 equal windows spanning the clip, then the 30 s windows inside the one that holds the evidence
Learning rate	1×10^{-4}, cosine without warmup
Items per optimizer step	8
Deployed checkpoint	step 2000
Answer adapter ϕ_{ans}	
Data	3,404 instances, gold segments from OmniVideo-100K (Cai et al., 2026), over 1,022 distinct gold source videos, plus 175 held-out instances
Instance construction	3 segments per instance (gold +2 distractors drawn from *other* recordings); the source question's 4 options
Target and loss masking	bare gold option letter plus the end-of-turn token; loss on those tokens only, per-item target-token mean
Learning rate	1×10^{-4}, cosine without warmup to zero at step 900
Items per optimizer step	12
Deployed checkpoint	step 900
Deployed inference	
Block length	$\Delta = 600$ s, non-overlapping
Localization-pass media	video at 0.5 fps capped at 24 frames; all block audio at native rate
Candidate windows	$\delta = 75$ s each
Selection	top $B = 3$ blocks $\times$ top $W = 3$ windows, so $\leq BW = 9$ windows
Answer-pass media	32 frames per window (up to 2 fps) plus that window's audio at native rate
Answer-pass transcript outline	question-ranked minute lines, at most 4,000 tokens
Answer-pass position cap	36,864 tokens
Decoding	greedy, at most 256 new tokens

cross-entropy of Eq. 3, summed with equal weight. Localization checkpoints are scored every 500 steps on a fixed 247-question set from the validation split by whether the evidence span's midpoint falls inside one of the three 30 s windows the localization pass shortlists on the two-level grid of Table 4, the best 30 s window inside each of its three highest-scoring first-level windows; step 2,000 scores best and is deployed. The answer adapter runs 900 steps, about three epochs over the instances.

Prompts and the letter read-out. The localization pass is one user turn with no system message: the block's video, the block's audio, then the text below, where the windows are listed in absolute clip time and the question is the benchmark's own stem; the question's answer options are not shown.

```
In which time window does the evidence to answer this
question appear?
{question}
A. 10:00–11:15
B. 11:15–12:30
...
H. 18:45–20:00
```

Table 5: **MiniCPM-o 4.5 and the causal-access protocol.**

MiniCPM-o 4.5	
Backbone	MiniCPM-o 4.5
Attachment points	attention query, key, value and output projections, and the MLP gate, up and down projections
Localization objective	overlap-fraction BCE over up to eight 75 s candidate windows
Localization supervision grid	one level for every clip
Localization items per step	12
Data-parallel ranks	6 for the localization adapter, 3 for the answer adapter
Answer data	13,192 instances, gold segments from OmniVideo-100K and LongVALE, over 4,215 distinct gold source videos
Answer deployed checkpoint	step 1100, where the cosine schedule ends
Localization-pass media	video at 1 fps capped at 320 frames per block
Answer-pass media	10 frames per window
Answer-pass transcript outline	none
Causal-access protocol on StreamArena	
Input	the causal prefix of the recording up to the query time
Answer weights	the base model, no adapter mounted
Answer-pass transcript outline	built from the text transcribed by the query time, question-ranked, at most 4,000 tokens
Reserved window	the 75 s ending at the query time, added to the retained windows (replacing the lowest-scoring one when the context is full) unless a retained window already overlaps it at an intersection-over-union of at least one half
Decoding	free text; the generation itself is the answer, with no letter parser

```
Answer with the option's letter from the given choices
directly.
```

The hidden state at the last prompt position is read against the output-head rows of the lettered options alone, each letter tokenized in its prompt context, and those logits are the $\ell_{m,k}$ of Eq. 1.

The answer pass is also one user turn: the transcript outline where the recording has speech, then a line saying several discontinuous segments follow, each labelled with its clip timestamp, to be reasoned over jointly, then the retained windows in time order, each introduced by `Segment` i `[`s_0`s–`s_1`s]:` carried as a video clip with its audio, then the question. On TraceAV the question closes with the benchmark's official prompt, which asks for the letter alone and for comma-joined letters when several options are correct. The reply is decoded greedily and parsed by its first standalone letter among the question's option letters; on TraceAV the final letter group is read as a set and must match the key exactly. A reply with no option letter is scored wrong.

Media context. Both passes cap every frame at 336^2 pixels, so a 16:9 source decodes at 448×252, which the backbone's processor raises to its minimum frame area at 512×288: 72 visual tokens per frame (144 per two-frame pair), the figure Appendix A.4 uses; audio costs 13 tokens per second at either stage. The localization pass samples 0.5 fps up to 24 frames, so a block of 48 s or more carries 24 frames and a shorter one 0.5 fps of it, never fewer than 4; a full block costs about 9,700 tokens, of which 7,800 are audio. The answer pass samples 2 fps up to 32 frames per window, so a window of 16 s or more carries 32 frames and a shorter one 2 fps of it, never fewer than 4; frame counts are even throughout. Audio is always the whole span at the native rate. A recording tiles into $\lceil T/600\,\mathrm{s}\rceil$ blocks and a block into windows, the last of each keeping the remainder, so a trailing window is at most 75 s and a trailing block of under 75 s holds one candidate.

The transcript channel's rendering. A candidate window is rendered as the transcript segments overlapping it, in time order, cut at 1,200 characters: segments are appended whole until the next would overflow, and the rest of the window is replaced by an ellipsis; a window with no speech reads "(no speech)". The block is rendered as one line per lettered window carrying its time range, followed by the window-selection prompt the media channel uses, with no media element in the conversation, and scored by the base model from the same option-letter logits.

The transcript outline. The transcript is produced once per recording by Whisper large-v3 under `faster-whisper`, greedy with voice-activity filtering and the language detected automatically, as segments carrying start and end times. Segments are bucketed into minutes by their start time and joined into one line per minute headed by that minute's range; a minute with no speech produces no line. To rank the minutes against a question, the stem and every option text are lower-cased and split into alphanumeric words. A minute scores the sum, over the query words it contains, of $\log(N/\mathrm{df})$, with N the number of spoken minutes of that recording and df the number of them containing the word, so a word spoken in every minute weighs nothing. An outline that fits the $4{,}000$-token cap is passed verbatim. Otherwise every minute header is kept, and the minutes with a positive score are admitted whole in descending score (the earlier minute first on ties) while each fits the space that remains. The space left after that is shared by the minutes not admitted: each keeps the same leading fraction of its own tokens, closed by an ellipsis, so no minute disappears. Tokens are counted by the backbone's tokenizer, lines are emitted in time order under a one-line outline header.

A.4 Cost accounting: a bounded peak footprint

Let n_q be the number of question and instruction tokens, n_ω the transcript outline's token cap, and n_δ the maximum token count of one selected window at the per-window context. Since at most BW windows enter the answer stage,

$$n_{\mathrm{ans}} \le n_q + n_\omega + BW n_\delta + n_{\mathrm{sep}}, \tag{5}$$

where n_{ans} is the answer-pass length in tokens and n_{sep} accounts for timestamps and separators. Every term on the right is a configured constant, with no term in T, and the localization-pass length is independently bounded by one compressed block.

For Qwen3-Omni, the terms of Eq. 5 translate to tokens as follows: one answer-pass frame costs 72 tokens at the deployed pixel cap (Appendix A.3) and audio ≈13 tokens per second, so one selected 75 s window carries $n_\delta \approx 3{,}300$ tokens and the nine-window answer pass ≈30,000, inside the deployed answer-pass cap of $36{,}864$ tokens, itself well under the backbone's $65{,}536$-token position limit. The transcript outline adds at most $4{,}000$ tokens, counted in every answer-pass figure below.

Query work is one localization pass per block plus the single bounded answer pass, so the pass count is fixed at $M+1$ (§3.2.2): median 13 passes on VideoOdyssey and single-digit medians elsewhere. In the deployed runs, every block pass stays under $10{,}000$ tokens on Qwen3-Omni and under $30{,}000$ on MiniCPM-o 4.5, and no answer pass reaches the deployed position cap. *Cumulative* prefill tokens are 2.2–$2.4\times$ those of a single whole-clip pass over the same questions (Qwen3-Omni medians on TraceAV, LVOmniBench and a stratified MMOU subset). *Working memory* is the peak of one pass: the weights plus the key–value state of at most one answer pass at the position cap, so **peak context and peak memory are bounded at any duration**.

The accounting is per-question worst case: the media prefix is reusable. The block tiling and each block's fixed grid of candidate windows are set by the clip alone, the backbone is frozen, and every localization-pass conversation places the block's media tokens *before* the question text. The media decode, the encoder features, and the media-prefix KV of every block pass are therefore identical by construction across questions about the same video, and the question suffix is ≈2% of a localization pass. These prefixes are precisely the artifacts the prefix-reuse literature caches and serves (Gim et al., 2024; Zheng et al., 2024; Liu et al., 2024), and that streaming video models keep across questions as the stream's KV cache (Di et al., 2025; Chen et al., 2026).

Paying for each block's media prefix once per *video*, with every block prefix of the video kept until its last question, takes the median per-question cumulative prefill on VideoOdyssey (10.6 questions per video) from 148.1k to 42.5k tokens ($0.29\times$). This reuse trades memory for prefill: the cache holds one localization-pass prefix per block, so it grows linearly with duration and lies outside the per-pass memory bound above, which the deployed runs meet with no cache. Only the localization pass reuses; the answer pass reads per-question windows. A vLLM serving run shows the reuse is realizable. We run the localization pass of six 1.5–2.1 h videos through a four-way tensor-parallel vLLM server whose key–value pool holds one video's block prefixes. Automatic prefix caching hits the cached block prefixes at 87.9% and cuts the median localization-pass wall-clock of a video's later questions by $1.6\times$ against the same server with caching disabled.

Question-in-the-loop re-observation (Pan et al., 2025; Wang et al., 2026b) decides *what media to process* from the question, so its media work re-runs per question by construction. Memory-bank streaming models (Song et al., 2024; He et al., 2024) also ingest once, into a lossy fixed-size state; LEAP's cached block prefixes are exact encodings whose size grows with the recording, and the selected windows are re-read at the per-window context.

A.5 Relation to prior search over long recordings

First, training matches inference in being $\mathcal{O}(1)$ context in duration: Equation 3 is defined within one block, so no training step reads past one block. Selection policies optimized by reinforcement from the answer (Pan et al., 2025; Li et al., 2026) instead carry a live answering model through every update. Frame-Voyager keeps that model offline, ranking frame combinations by its loss once and training the selector on the frozen labels, but its selector still reads candidate frames drawn from across the whole recording in one context (Yu et al., 2025). Second, the supervision never sees answer correctness, whereas a selector trained on answer reward learns which inputs make one particular answering model produce a response that is marked correct.

Cross-block ranking needs no training of its own: every pass shares one prompt and one adapter, and the blocks are ranked on those logits directly (Appendix C.6). The scorer is the answering backbone itself: relevance is an option-letter logit read from the frozen backbone under the localization adapter. No separate retriever, captioner or embedding index stands beside the model, and ranking reuses logits the localization pass has already produced: ranking adds no parameters, no pass, and no term in B on top of the $M+1$ cost. The score is question-conditioned and jointly audio-visual with full-rate audio, where previous selectors score one modality (Wang et al., 2026b; Shao et al., 2026), reach the audio through the visual track (Diao et al., 2025), or return a single interval (Zhang et al., 2026a).

Finally, no pass over a recording longer than one block holds the whole recording in context: two-stage systems process the full recording in their first stage (Hannan et al., 2025; Li et al., 2026) and agentic methods re-observe the source per question (Tao et al., 2025; Xing et al., 2026; Zhang et al., 2026b); in our method every pass reads one block or a bounded set of retained windows.

LEAP differs from Zoom-Zero's divide-and-conquer variant (Shen et al., 2025), which also splits a recording into non-overlapping windows scanned independently, in two ways. First, our window score is an option-letter logit the localization pass already produced, so ranking adds no pass; its score is a trial answer's confidence, which costs a second pass per block and varies mostly between questions, the part of a score a within-question ranking cannot use (Appendix C.6). Second, our scan is jointly audio-visual, reading each block's audio at the native rate, while its scan is visual.

B Evaluation Protocol and Baselines

B.1 Systems, contexts, and reporting conventions

The baselines. Table 6 lists the single-pass baselines on Qwen3-Omni-30B and what each reads. The *context-filled montage* stands in for the whole clip wherever one pass has to fit a recording of any length: 32 frames spread over the recording plus uniform 10 s audio slices, packed until the encoded input, media and question together, reaches 60,000 tokens, leaving room for the reply under the 65,536-token position limit. A short recording therefore keeps nearly all of its audio and a four-hour one keeps a fraction. On VideoOdyssey the official recipe is our approximation of the benchmark's 64-frame and audio-montage pipeline (Appendix B.3).

The whole-clip limit. Many of these videos are too long for a single whole-clip pass: some TraceAV and LVOmniBench questions, and the large majority of VideoOdyssey questions, exceed the whole-clip limit, the longest recording that fits one whole-clip pass inside the 65,536-token position limit, roughly 81 minutes at the media context used here (32 frames plus full audio). OmniVideoBench and MMOU are the only benchmarks of §4 entirely within it. LEAP's answer pass runs under a tighter cap of 36,864 tokens (Table 4), which its nine windows never reach (Appendix A.4).

Reporting conventions. Percentage-point differences are written *pp*. Every paired difference in this paper is read on one inferential quantity: the two-sided 95% bootstrap interval on the paired

Table 6: **The single-pass baselines on Qwen3-Omni-30B.** Each reads the whole recording, or on StreamArena the causal prefix, in one answer pass; *uniform windows* instead keeps the answer pass of LEAP and places its windows at equal spacing. *Frames* are spread evenly over what the baseline reads; *full track* = the native waveform of that span; *official montage* = 64 ten-second audio slices at evenly spaced starts, concatenated into one track, as cut by the benchmark's released evaluation code. The official recipes of TraceAV and LVOmniBench also carry the benchmark's own prompt.

Baseline	Frames	Audio	Read in
Official recipe[‡]			
TraceAV	up to 256	full track	Table 1, Figure 2a
LVOmniBench	128 at 336^2	full track	Table 1, Figure 2a
VideoOdyssey	64	official montage	Tables 1 and 20, Figure 2a
MMOU	32	full track	Table 1, Figure 2a
Frame-matched whole clip	288	full track	Tables 1 and 7, Appendix B.4
Whole clip	32	full track (none on LVBench and Video-MME in Figure 2a)	Tables 1 and 8, Figures 2a, 4, 13 and 16
Context-filled montage	32	10 s slices up to 60,000 tokens	Table 7, Figures 4, 13 and 15
Native-sampling whole clip	2 fps, at most 128	official montage, or none	Table 18, Figures 2a and 14
Whole prefix	32	full track	Table 9, Figure 2a
Uniform windows	32 per window, up to 9 windows	each window's own	Table 9

difference, resampled at the source video, since many questions share one recording, at a fixed seed and at least 10,000 resamples. An effect is *significant* exactly when that interval excludes zero. The error bars on the length curves of Figures 4 and 13 are per-line intervals resampled at the source video in the same way, and carry no paired claim. Unless a caption says otherwise, every denominator is the full official question set, with failures scored as errors where gold labels are public.

B.2 Concurrent systems on these benchmarks

Video-SALMONN-R[3] (Li et al., 2026) trains a two-pass re-watch policy by reinforcement learning on a Qwen3-VL backbone with a separate audio encoder, and OmniRAG-Agent (Zhu et al., 2026) reports OmniVideoBench on a self-selected subset; its reinforcement-trained version exists only on smaller backbones, with no released weights, so we run its untrained version on our stack (Table 7). Agentic pipelines (Tao et al., 2025) run frontier-API components once per question, and offline index-and-search systems (Yin et al., 2026) add a per-video indexing cost. OmniVideo-100K (Cai et al., 2026) also fine-tunes omni backbones on cross-segment audio-visual evidence chains, but as whole-clip instruction data with no retrieval at inference. LEAP's answer adapter trains on instances from its gold segments (Table 4) and reads the windows a trained selector retrieves (§3.4).

OmniVideo-100K's released 30B checkpoint is our own backbone fine-tuned that way, so we run it on our stack under our own protocol (Table 7). LEAP leads it on every benchmark except MMOU. On MMOU, LEAP's retrieval on that checkpoint lifts it significantly, from 70.1 to 71.6: the checkpoint answers the windows LEAP selects, with LEAP's transcript outline, instead of the whole clip.

The OmniRAG-Agent loop wraps the same unmodified backbone in an external index: the recording is time-compressed to a fixed-length clip placed in the first turn, and any evidence beyond that clip arrives as CLIP-ranked frames and speech-transcript segments fetched over a bounded sequence of tool calls. LEAP outperforms OmniRAG-Agent on all four benchmarks.

Table 7: **Concurrent systems on the same backbone.** Accuracy (%). The first two rows read the same input: the whole recording sampled uniformly and frame-matched to the most frames our answer pass reads, with its full audio, or on VideoOdyssey the context-filled audio montage. *OmniVideo-100K SFT* = the released checkpoint fine-tuned on that instruction set, no retrieval stage (Cai et al., 2026). *OmniRAG-Agent loop* = the unmodified backbone inside that paper's released retrieval agent, scored by our answer extractor (Zhu et al., 2026). * = the benchmark's own runs, as TraceAV reports them (Feng et al., 2026): the bare backbone, and two untrained agents driving it, *AVP* (Wang et al., 2026b) and *Audio-Guided* OmniAgent (Tao et al., 2025). — = no number. TraceAV columns follow its two published tables: subtask macro-averages over the general and over the hallucination sub-tasks. **Bold** = best, underline = second best in each column.

System	TraceAV General	TraceAV Hallucination	LVOmni	VideoOdyssey (1–4h)	MMOU
Qwen3-Omni-30B	<u>53.6</u>	68.0	<u>40.4</u>	37.9	57.8
+ OmniVideo-100K SFT	52.4	65.0	39.3	<u>43.1</u>	**70.1**
+ OmniRAG-Agent loop	39.5	63.4	27.5	26.6	34.8
Qwen3-Omni-30B, as published*	48.4	67.5	—	—	—
+ AVP*	47.1	<u>69.8</u>	—	—	—
+ Audio-Guided*	42.4	57.9	—	—	—
LEAP	**60.9**	**76.2**	**45.8**	**53.7**	<u>66.3</u>

TraceAV's own appendix runs two further untrained agents on this backbone, AVP (Wang et al., 2026b) and Audio-Guided OmniAgent (Tao et al., 2025), with the backbone standing in for the models each was released with (Table 7). LEAP leads both agents on both of its tables.

B.3 The same-stack official-recipe baselines

Where the published same-backbone number was produced on a pipeline we cannot reproduce bit-for-bit, we instead re-run the official *input recipe* on our own stack (same backbone weights, same harness) and read method gain only against that same-stack baseline. These are the ‡ cells of Table 1. Paired against them on the same questions, LEAP without answer training leads significantly on TraceAV, VideoOdyssey and MMOU. On VideoOdyssey that baseline re-implements the official audio-montage API pipeline on our stack. With that montage (Table 6) the model hears about eleven minutes of any recording. Our re-implementation departs from the official pipeline in three ways.

Audio. The model hears the complete montage waveform. The official pipeline base64-encodes a 16 kbps MP3 of the ≈640s montage into an OpenAI-compatible `input_audio` field.

Frames. We pass the 64 frames as a *video* element, so the model sees temporal order and adjacent-frame pairing. The official pipeline sends the same 64 frames as 64 independent still images side by side, with no timestamps and no video-side temporal encoding.

Denominator. The official protocol moves unanswered or unparsed questions *out* of the denominator, so the published figure is accuracy over the answerable subset only.

Smaller divergences remain: a forced-guess instruction in the official prompt, message ordering, decoding temperature, and JPEG/MP3 lossy round-trips.

The official recipes of LVOmniBench and MMOU. LVOmniBench's official pipeline runs on HF transformers with greedy decoding and the full native waveform, with more frames than ours. It still needs a same-stack baseline, on two counts. **The official recipe does not fit the backbone it is run on**: the official model wrapper pins neither the frame count nor the per-frame resolution, so both fall to the video-loading library's defaults, sized for a 128k-token context while this backbone's position limit is 65,536. Re-encoding the official conversation verbatim puts every question past the limit: the 768-frame video stack alone exceeds it before a single audio token is added. **And the**

Table 8: **Input-form control.** LEAP and the whole clip carry the same answer LoRA and no transcript outline, and differ in input form: the whole clip is read up to the backbone's position limit, the selected windows at the answer-pass cap. Under *full denominator* a clip too long to encode whole inside the backbone's position limit is scored wrong for the whole clip; *both forms fit* drops those questions. TraceAV = accuracy over all its questions, hallucination sub-tasks included. VideoOdyssey is absent: most of its recordings are too long to encode whole.

	Full denominator		Both forms fit	
Benchmark	Whole clip	Selected windows	Whole clip	Selected windows
TraceAV	57.3	60.4	59.9	60.6
LVOmni	42.9	47.5	43.7	47.3
OmniVideoBench	42.3	43.3	42.3	43.3
MMOU	61.8	65.2	61.8	65.2

official harness drops failures from the denominator, filtering out every question whose pass failed (out-of-memory included) or whose answer failed to parse, with no skip count reported.

An official-style base whole-clip read on our stack (official prompt, full native waveform, greedy, no adapter, 128 frames at 336^2) is LVOmniBench's baseline in §4.

MMOU publishes a scoring protocol but no input recipe, so its baseline is an official-style whole-clip read on the base model at our media context, greedy, one pass in the original option order, whereas the official scoring uses five option permutations, majority voting, and best-of-prompt selection.

B.4 The frame-matched whole clip

Table 8 pairs LEAP against the whole clip at our own media context of 32 frames plus the full audio. A second control matches the *frames* instead: the whole clip answered at LEAP's own ceiling of 288 frames (9 windows $\times$ 32), sampled evenly across the recording and carrying the full native waveform, and answered by the base model. Its audio contains everything LEAP's does, since LEAP's windows carry no more audio than the whole track. At 288 frames some clips no longer fit: questions on TraceAV, LVOmniBench and MMOU overflow and are scored wrong.

It is read against LEAP without answer training. On the full denominator, where a question the frame-matched whole clip cannot fit is scored wrong for it, LEAP without answer training leads significantly on TraceAV and MMOU, and on MMOU also once those questions are dropped. LEAP itself leads significantly on TraceAV, LVOmniBench and MMOU on both denominators.

B.5 The causal-access protocol on StreamArena

The benchmark. StreamArena (Zhang et al., 2026c) poses open-ended questions at timestamped moments of hour-long recordings and scores a free-text answer by whether it contains the reference answer's factual core. We report its historical-retrospection task, every question answered independently from the recording alone. Every pass's input ends at the query time, so the 600 s block grid tiles exactly the causal prefix, a median of 4 and at most 12 blocks per question.

The same-stack comparison on Qwen3-Omni-30B. LEAP and the whole-prefix baseline share the frozen backbone, the prompt (stating the query time and that *now* is the last segment's end) and the judge, a local Qwen3.6-27B applying the benchmark's binary criterion with one vote.

LEAP here is the deployed pipeline with no streaming training (Figure 5): the localization adapter scans the causal blocks at the localization-pass context, retains the top 3 and shortlists 3 windows each, and the base model answers the retained windows at 32 frames per 75 s window with the windows' own audio. The window ending at the query time, or a retained window overlapping it at an intersection-over-union of at least one half, is always among them (Table 5).

The whole-prefix baseline answers the whole causal prefix in one pass (the whole-clip recipe: 32 frames spread over the prefix with its full audio); questions whose prefix does not fit are scored as errors. Restricting the pairing to the questions whose prefix fits leaves the lead significant (+5.60 pp).

Table 9: **Historical retrospection under causal access on StreamArena.** *HR* bins = minutes from the query time back to the evidence, under the benchmark's own bin names. *Query-time input* = what the model reads when the question arrives. Published rows are the benchmark's own table under its judge: ° = closed source, [a] = marked there as an author-finetuned backbone. *LEAP – media* scans the media blocks with the localization adapter mounted, *Base selector* the same blocks without it, and *LEAP – transcript* the transcript index built as the stream arrives; all three answer the retained windows from the raw media. *Transcript only* answers the windows of *LEAP – transcript* from their transcript instead. **Bold** = best, underline = second best in each column within a group.

				evidence distance (min)			
				L1	L2	L3	L4
System	Backbone	Query-time input	HR avg	≤5	5–15	15–30	>30
Published, benchmark judge							
StreamMind	Qwen3.5-397B-A17B	continuous memory	34.9	**31.5**	**46.7**	**34.6**	17.1
Qwen3.5-Omni°	—	whole prefix	**35.8**	**31.5**	44.8	34.1	**25.4**
AURA[a]	Qwen3-VL-8B	recent window	22.7	25.4	27.0	24.3	10.5
MiniCPM-o-4.5[a]	Qwen3-8B	recent window	9.8	10.5	13.3	7.6	5.0
StreamForest	Qwen2-7B	whole prefix, compressed	14.4	19.3	13.6	14.1	11.0
Same stack, local judge, no streaming training							
Whole prefix	Qwen3-Omni-30B	the whole prefix, one pass	25.4	22.1	34.2	25.5	12.6
Uniform windows	Qwen3-Omni-30B	≤ 9 uniform windows	20.2	23.2	25.8	17.9	9.3
LEAP – media	Qwen3-Omni-30B	≤ 9 windows, media scan	**32.0**	**26.5**	**41.2**	29.3	23.6
Base selector	Qwen3-Omni-30B	≤ 9 windows, media scan	27.1	23.8	32.1	27.2	21.4
LEAP – transcript	Qwen3-Omni-30B	≤ 9 windows, transcript scan	30.1	24.3	36.1	**30.4**	**24.7**
Transcript only	Qwen3-Omni-30B	same windows, as text	22.3	14.4	27.9	22.8	19.8
Native streaming	MiniCPM-o 4.5	last 30 s	8.6	13.3	7.3	4.9	9.9
LEAP	MiniCPM-o 4.5	≤ 9 selected windows	15.1	18.2	18.8	10.3	9.9

The uniform-windows baseline keeps LEAP's context (up to 9 windows of 75 s) and replaces the ranking by windows spread equally over the causal prefix, the last one ending at the query time; a prefix that fits that context is read whole. It reads at least as many windows as LEAP, and more on nearly every question where LEAP under-fills its context, and is answered by the same weights, prompt and judge. LEAP's ranked windows cover the annotated evidence on nearly three quarters of the historical-retrospection questions, against about half for the uniform windows.

A transcript index built as the stream arrives. The index is built in time order (Figure 5). faster-whisper large-v3 transcribes the audio in 30 s chunks with no lookahead and no overlap, each chunk decoded with only the previous chunk's text as context. Every transcribed segment is stamped with the moment its text became available under a real-time schedule: one ASR worker per stream, a chunk decoded once its audio is complete. A question reads only the segments available at its query time. The base selector scores every causal block's candidate windows from that text, as the transcript channel does offline, and its retained windows go to the same answer pass, the base model, with the same reserved window for the most recent 75 s. On one L40S the transcription runs at a real-time factor of 0.028, and text arrives a median 0.8 s after the chunk's audio ends (1.3 s at the 95th percentile). The index occupies 70 KB per recording hour, beside 0.88 GB of raw archive.

Its windows cover the annotated evidence on 70.9% of the historical-retrospection questions, against 72.1% for the media scan of LEAP. With the evidence modality labeled by the judge model from each question and its annotated evidence description, the transcript channel covers spoken evidence significantly more often than that media scan and visual evidence slightly less often.

The two scans differ in the selector as well as in the index, so a further comparison holds the selector fixed: the base selector scans the media with the localization adapter unmounted, and its windows go to the same answer pass (*Base selector* in Table 9). Under that one selector the transcript channel covers the evidence significantly more often than the media, by 20.6 pp, in every distance bin and for visual as well as spoken evidence, and its answers are significantly more often correct. Without the localization adapter the windows cover about as much of the evidence as the uniform ones; block

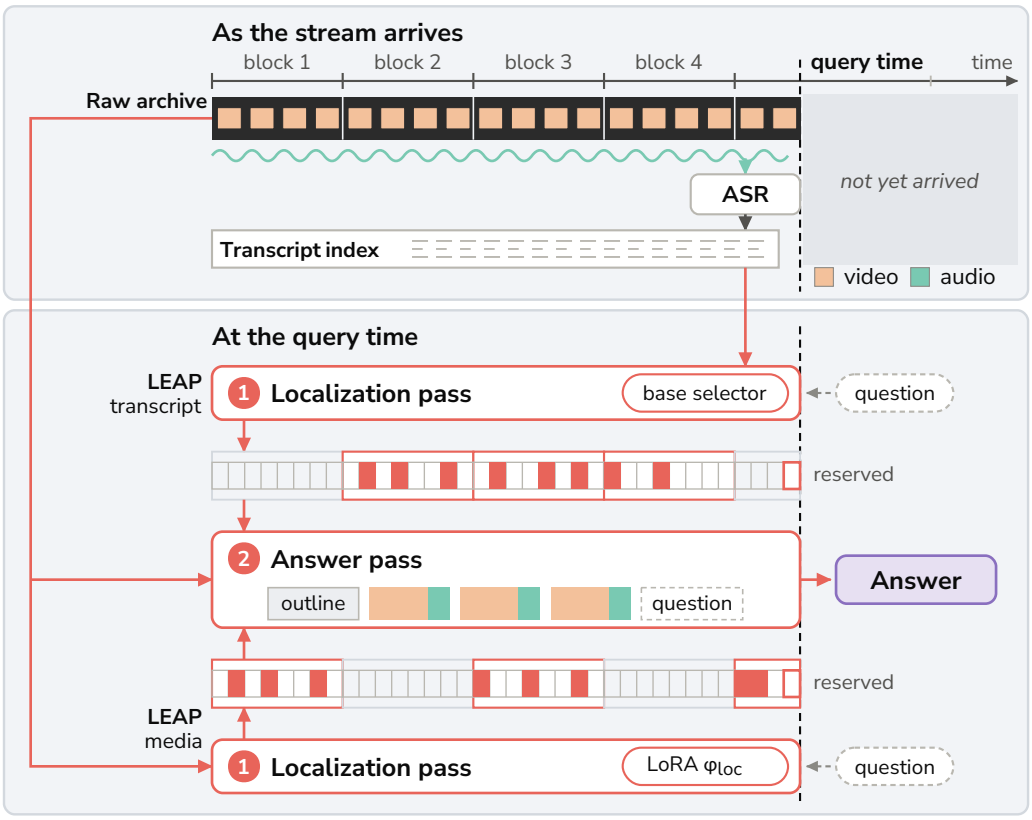

Figure 5: **LEAP in the streaming scenario.** One time axis, cut by the query time; nothing to its right exists when the question is asked. *Top*: the raw archive and the transcript index built beside it. *Bottom*: the two selectors of LEAP and the one answer pass they share, which reads the retained windows from the archive and the outline from the index. Faint cells are candidate windows, solid cells the windows retained inside the blocks each selector keeps, and the outlined cell the slot held for the moment before the query, whether or not the ranking kept its block.

coverage stays close to the adapter's, and the gap is in the choice of window inside a retained block. Mounting the localization adapter closes that gap, raising media coverage by 21.8 pp.

Answering the retained windows from the raw media leads answering from their transcript alone significantly in all but the farthest distance bin, and leads the uniform windows (Table 9). Heard and not seen (each retained window as its audio alone), the same windows recover 3.5 pp of the gain over the transcript; the frames on top of the audio add the remaining 4.2 pp, and both parts are significant.

The transcript outline at the answer pass. LEAP's answer pass reads the transcript outline of §4.5, cut at the query time. Without it, LEAP's accuracy barely moves overall.

Re-reading the selected windows at the published resolution and frame rate. The published offline rows read up to 128 frames at up to 720p and the streaming rows one or two frames per second. Re-reading the selected windows of LEAP at 1280×720 with 12 frames per 75 s window, trading frames for resolution, or at 80 frames per window (1.07 fps) at the deployed resolution barely moves historical-retrospection accuracy against the deployed per-window context over the same windows, all three read without the outline. Both re-reads run under a position cap widened to 65,536 tokens, since the denser windows overflow the deployed cap. Cutting the answer pass to the localization pass's own frames, about three per window with the audio unchanged, costs 2.5 pp significantly.

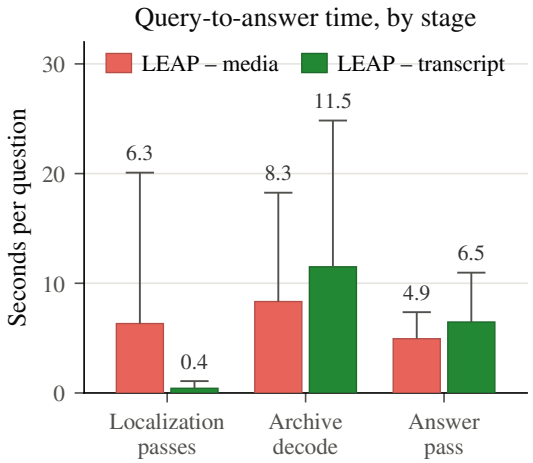

Figure 6: **Query-to-answer time by stage on StreamArena.** Seconds per historical-retrospection question on one H100, one group per pipeline stage: the localization passes over the causal blocks, decoding the retained windows from the archive, and the answer pass. The bar is the per-question median and the whisker above it reaches the 95th percentile.

Latency. Figure 6 splits query-to-answer time on a single H100, with no prefix caching and no batching, into every causal block's localization pass, the media decoding of the retained windows, and the one answer pass. Under the transcript index nearly all of the localization time disappears, and decoding and answering remain. Localization's tail grows with the number of causal blocks before the query, and the transcript index removes that tail along with the median.

The same-stack comparison on MiniCPM-o 4.5. The smaller backbone answers this benchmark at the media context of Table 5 and with no streaming training, and it supplies a baseline the larger one cannot, its native streaming mode: half-duplex over the 30 s before the query time, one frame paired with one second of 16 kHz audio per second. Both conditions answer with the released weights (LEAP's selector carries the localization adapter of Table 5), greedy, one local judge, every question answered alone, over the same questions and clustered interval as the Qwen3-Omni-30B group. LEAP leads the native streaming mode inside the same-stack group (Table 9); the published rows are scored under the benchmark's judge with the earlier turns of its multi-turn protocol in context.

B.6 Contamination

The two source corpora of the deployed training pools (LongVALE and OmniVideo-100K, 12,454 videos, a superset of every video behind either adapter's training items) were audited against every benchmark but StreamArena at three levels where the source media are available: video identifier; perceptual frame hash on a ≈5 s grid, the 300 pairs per benchmark with the most offset-consistent matches re-scored by grayscale normalized cross-correlation; and Haitsma–Kalker audio fingerprints matched under offset consistency. On TraceAV, whose recordings mostly carry a YouTube identifier, the frame and audio levels cover the recordings without one.

At the identifier level, no benchmark video meets a training set of the backbone it is reported for.

C The Retrieval Stage

Table 10 removes LEAP's components one at a time. Reading the whole clip in place of the retrieved windows lowers accuracy significantly on TraceAV and VideoOdyssey. The windows LEAP's localization LoRA selects are answered significantly better than those of the base selector on LVOmniBench, VideoOdyssey and MMOU. LEAP gains significantly from the transcript outline on TraceAV, VideoOdyssey and MMOU.

Table 10: **Removing LEAP's components.** Accuracy (%) on Qwen3-Omni-30B. Each row removes one component **from the row above**, and the answer LoRA answers in every row. − *transcript outline* answers the same windows without the outline; − *localization LoRA* lets the bare backbone's coarse scan pick the windows; − *retrieval* reads the whole clip in one pass (on VideoOdyssey, its audio-montage recipe). TraceAV = the mean over its twelve general sub-tasks. **Bold** = best, underline = second best in each column.

System	TraceAV	LVOmni	VideoOdyssey	MMOU
LEAP	**60.9**	45.8	**53.7**	**66.3**
− transcript outline	57.8	**47.5**	51.2	65.2
− localization LoRA	56.2	43.5	46.7	62.1
− retrieval	54.8	42.9	41.9	61.8

Table 11: **An off-the-shelf retriever as the ranking score.** Every row keeps the same candidate-window grid, block and window counts and answers with the localization adapter mounted; rows differ only in the score a candidate is ranked by. Under the trial-answer score the windows inside a block are still ranked by the base selector. All rows, the base selector and the localization adapter included, come from localization passes that split a recording's partial tail block differently from the deployed ones, so they compare only with each other. TraceAV = accuracy over all its questions, hallucination sub-tasks included. **Bold** = best, underline = second best in each column.

	Accuracy (%)				Final-window coverage (%)		
Ranking score	TraceAV	LVOmni	VideoOdyssey	MMOU	TraceAV	VideoOdyssey	MMOU
Base selector	58.3	39.0	41.6	56.2	92.0	57.3	82.7
Trial-answer confidence	58.8	39.0	41.5	56.3	90.2	48.3	82.7
SigLIP, frames	58.8	43.0	42.0	57.7	**96.2**	65.8	88.9
bge-m3, transcript	**60.8**	42.2	43.6	58.1	96.0	58.9	88.3
Localization adapter	59.2	**43.3**	**45.6**	**61.1**	**96.2**	**71.7**	**94.8**

C.1 Selection interventions: the retained-block count

Unless a paragraph says otherwise, every ablation in this appendix re-ranks or re-answers the stored per-window logits of the deployed localization passes. A paragraph that runs a new localization pass (another selector, a silenced or single-channel input, a retrained adapter) says so where it appears. The trial-answer variance of §C.6 reads a separate TraceAV localization pass, the 30 *s-window scan*: the same 600 s blocks with 30 s candidate windows, under a prompt that also offered a 'no evidence' option. The CG-Bench panel of Figure 7 and the CG-Bench window-count ablation of §C.4 read a CG-Bench localization pass under that same prompt with the 'no evidence' option. Table 11 reads localization passes that split a recording's partial tail block differently: a tail block of up to 240 s is one candidate window, a longer one eight equal windows. The two selector variants of Figure 3 share the localization pass, window enumeration and block ranking, so they differ only in the localization adapter. The answer passes of Figures 3 and 10, of Table 11, of the CG-Bench panel of Figure 7 and of the window-count and window-width ablations of §C.4 read no transcript outline.

An off-the-shelf retriever as the ranking score. A third selector fits the same comparison, with its own base-selector and localization-adapter references (Table 11). Keep the enumeration, the three-blocks-by-three-windows rule and the answerer, and rank windows by a retriever's similarity to the question: SigLIP image–text cosine over the window's frames, or bge-m3 text cosine against the window's speech transcript, the stem being the query in both (SigLIP truncates it to 64 tokens).

Pooled over four benchmarks, both retrievers answer more accurately than the base selector, and the trained adapter is significantly ahead of the frame retriever and of the trial-answer score (§C.6). Its lead sits where localization training buys the most evidence coverage: on VideoOdyssey and MMOU it covers more annotated evidence than both retrievers and answers more accurately, significantly against both on MMOU and against the frame retriever alone on VideoOdyssey.

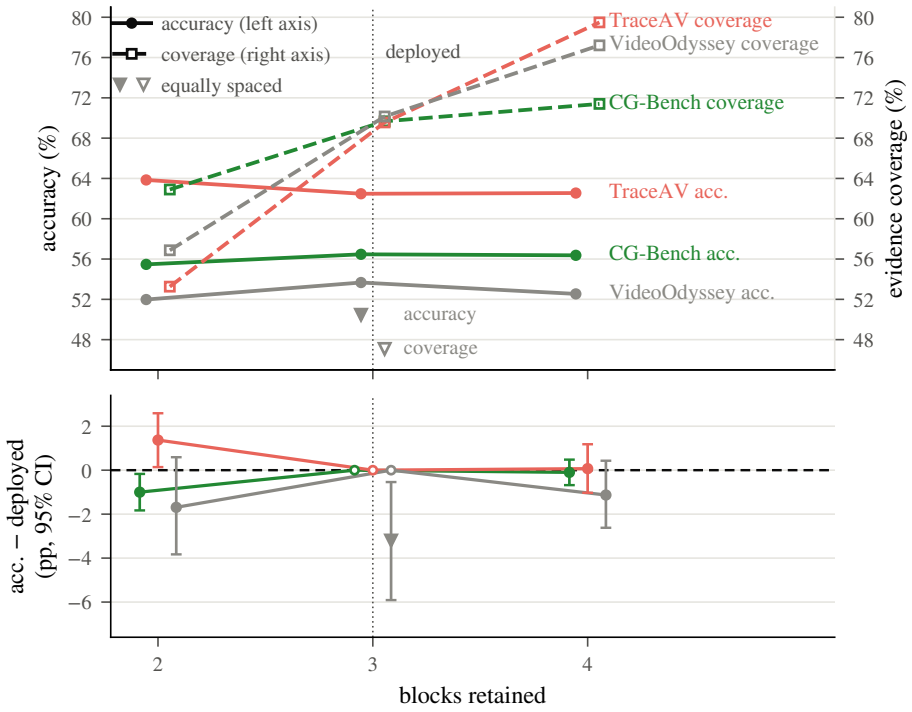

Figure 7: **Accuracy and evidence coverage against the retained-block count.** *Top*: levels, accuracy on the left axis and evidence coverage on the right; triangles hold the deployed count and replace the block ranking by equally spaced blocks. *Bottom*: the paired accuracy difference against the deployed count, with its bootstrap interval; the deployed count is drawn open at zero. TraceAV runs only on the recordings that tile into more blocks than the deployed count keeps.

Table 12: **What each channel's selected windows retain on TraceAV.** *Content-free uniform* spaces blocks and windows evenly and runs no pass. *Media decoded* = what a scoring pass decodes at query time; *tokens / block* = its measured length. *Recall* = the share of annotated evidence spans the selected windows deliver, over all questions and over those whose recording tiles into more blocks than the retained-block count (*ranking decides*). **Bold** = best, underline = second best in each recall column.

			Recall	
Channel	Media decoded	Tokens / block	All questions	Ranking decides
Content-free uniform	none	—	69.6	61.2
Listen-only channel	audio	6,855	71.7	67.0
Media channel	frames + audio	8,546	74.1	69.6
Transcript channel	none	1,579	**74.3**	**70.0**

What the block count buys. Evidence coverage rises with every block added, but accuracy does not follow it. With two blocks, LEAP answers significantly less accurately on CG-Bench and significantly more accurately on TraceAV. The block count is one setting for every benchmark, capped by what one answer pass holds. What still moves the answer at the deployed count is which blocks the ranking keeps: on VideoOdyssey, equally spaced blocks cost coverage and accuracy significantly (§C.6).

C.2 The transcript channel: what it retains and what it answers

What the selected windows retain. Table 12 reads TraceAV, whose questions are posed about audio and video together. Where the ranking decides (on recordings tiling into more blocks than the retained-block count, so block ranking has a choice), the transcript channel does not fall behind the media channel in mean per-question recall by more than a 2.5 pp non-inferiority margin. Neither the

Table 13: **The transcript outline at the answer pass.** Accuracy (%). Every row answers the same retrieved windows, grouped by system. *Question-ranked* (deployed) fills the outline question-first; *Uniform* thins every minute equally; *None* removes the outline. **Bold** = best, underline = second best within each group. TraceAV = accuracy over all its questions, hallucination sub-tasks included.

System	Transcript outline strategy	TraceAV	LVOmni	VideoOdyssey	MMOU
LEAP without answer training	Question-ranked	**60.8**	**43.0**	**46.7**	**61.4**
	Uniform	59.8	42.8	45.0	61.0
	None	58.2	42.7	43.6	61.1
LEAP	Question-ranked	**63.6**	45.8	**53.7**	**66.3**
	Uniform	63.2	46.4	50.7	66.2
	None	60.4	**47.5**	51.2	65.2

Table 14: **Where the outline's gain lands.** Accuracy (%) of LEAP, with the transcript outline present or removed. MMOU questions are grouped by how much of their annotated evidence interval the retained windows overlap; TraceAV questions by their annotated class.

Group	No outline	+ transcript outline
MMOU: overlap with the annotated evidence interval		
Evidence uncovered	48.0	55.9
Partially covered	63.8	65.5
Covered (at least half)	66.6	67.2
TraceAV: annotated question class		
Spatiotemporal localization	30.8	50.7
All other classes	63.8	65.1

content-free uniform selection (equally spaced blocks and windows, no pass run) nor the listen-only channel (the localization pass fed audio alone) meets that margin. Within TraceAV the two channels separate along how much of the evidence is spoken. Each question's recall difference, transcript minus media, regressed on the speech share of its annotated evidence span rises by $+11.68$pp $[+5.80, +17.41]$ from evidence with no speech to evidence entirely spoken.

The same comparison on CG-Bench mini, a video benchmark whose human-labeled clue intervals mark visual evidence, runs the other way. Over all its questions the transcript channel retains significantly less of those intervals than the media channel, -8.40pp $[-10.54, -6.31]$, and the deployed answer pass reading its windows answers significantly less accurately, -1.80pp $[-2.97, -0.64]$.

The seconds of Table 2. The per-question seconds are a sampled replay on one idle H100, one question at a time, all media decoded fresh, at the deployed answer pass's eight-worker decode concurrency; the transcript is that of Appendix A.3, rendered per window under its character cap.

C.3 The transcript outline at the answer pass

Where the outline gains. Table 14 groups the questions by whether retrieval already delivered the evidence. On MMOU each question's annotated evidence interval, where the annotation gives a usable one, is intersected with the windows the deployed selection retained; the annotation is used only to group the answer pass and never reaches a prompt. The gain on the uncovered group is significant, from a starting accuracy far below the benchmark's overall accuracy, while the covered majority barely moves. On TraceAV the split is taken along its question classes instead, and the gain concentrates in the spatiotemporal-localization class.

Filling the outline question-first. The deployed rule hands the outline's tokens to the minutes that score highest against the question; the *Uniform* row of Table 13 instead thins every minute by the same fraction. A question that matches nothing, or an outline the cap never truncates, renders the same prompt under both strategies, so a benchmark-level row is diluted by however often the outline already fits. Read on the truncated questions alone, the *Uniform* rule costs 3.57 points on

VideoOdyssey under LEAP and 1.81 points on MMOU under LEAP without answer training, both significant. They are also the cells that stay significant at the benchmark level.

C.4 THE WINDOW GRID

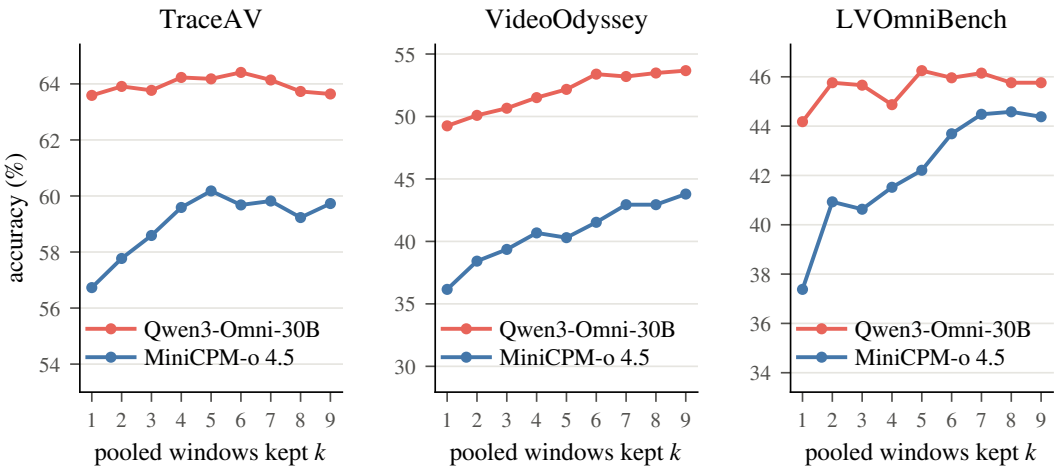

Figure 8: **Accuracy against the answer-pass window count.** The pooled windows are truncated to the top-k by window score. The deployed configuration keeps the whole pool, each curve's last point.

Where the deployed pool sits on the sweep. A sensitivity sweep re-answers the deployed selection with the pooled window set truncated to its top-k windows by window score, k=1..9, everything else held (Figure 8). The ranking is global over the retained blocks, so the axis is the answer pass's total window count, with the per-block shortlist W held at its deployed value throughout. It runs on both backbones on TraceAV, VideoOdyssey and LVOmniBench. On Qwen3-Omni-30B the TraceAV and LVOmniBench curves barely move from the first window on, and the VideoOdyssey curve rises until $k\approx6$ and barely moves after, every truncation to four or fewer windows reading significantly lower. On the smaller backbone the TraceAV curve reads significantly lower only at one and two windows; on VideoOdyssey its curve instead does not saturate, and every truncation of the pool reads lower, significantly so through the mid range. Its LVOmniBench curve rises until $k\approx6$, every truncation to four or fewer windows reading significantly lower.

Where the grid comes from. The block length is set by the longest training clip: the localization adapter trains on clips that fit inside a single block (Appendix A.2), and the longest clip in the LongVALE-derived corpus runs 597.9 s. At any larger Δ no training clip fills a block. With Δ fixed by the supervision, three axes stay free: the window width δ, which sets $K=\lceil\Delta/\delta\rceil$ (8 at δ=75 s); the per-block shortlist W; and the retained-block count B.

The deployed grid survives a two-benchmark ablation. We ablate the two axes Figure 7 leaves, W and the window width, on CG-Bench mini and VideoOdyssey. *Window count:* re-pooling the identical stored localization passes at W=2 barely moves accuracy on either benchmark, even though the dropped third window measurably costs clue-interval coverage on CG-Bench. *Window width:* doubling the window to 150 s (at the same W=2, since three 150 s windows per retained block would overflow the answer-pass position cap) barely moves VideoOdyssey and drops significantly on CG-Bench. On CG-Bench both widths use a localization adapter retrained with single-level candidates at that width, so width is the only difference there; on VideoOdyssey the 75 s variant keeps the deployed adapter. The human clue intervals are consistent with dilution: the wider windows overlap the human clue intervals *more* at the matched window count, yet answer less accurately.

Both counts are capped from above by one answer pass. Adding a fourth block overflows one answer pass: a 4-block configuration pools up to 12 windows of 75 s, whose combined length reaches 36.9–45.3k tokens on VideoOdyssey's overflowing questions against the deployed position cap of

Table 15: **The block-ranking score, ablated by component.** Each row re-ranks the same localization passes on VideoOdyssey. The deployed score is a monotone map of the block's average logit plus its best window's margin above it; *Deletes* = the component a variant removes, a dash neither. A *partial block* is a clip's tail block, with fewer candidate windows; *partial block retained* = the share of questions retaining it. **Bold** = best, underline = second best in each coverage column.

		Evidence coverage (%)		Partial block
Ranking score	Deletes	all questions	no partial block	retained (%)
$g_m = \max_k \sigma(\ell_{m,k})$ (deployed)	—	<u>70.2</u>	**78.7**	67.6
$\max_k \mathrm{softmax}(\ell)_k$	average, renormalized	63.8	73.8	62.5
$\max_k \ell_{m,k} - \bar{\ell}_m$	average, subtracted	**72.9**	<u>78.0</u>	11.8
$\bar{\ell}_m$	margin	33.8	39.0	82.1
$\frac{1}{K_m}\sum_k \sigma(\ell_{m,k})$	margin; mean of σ	23.8	26.8	77.8

36,864. The add-a-block variants of Figure 7 are answered with the cap widened to 49,152 for exactly the overflowing window sets. $W{=}4$ pools as many as twelve windows and meets the same cap.

C.5 Where the windows sit: the random-placement control

A control condition bounds the selection stage from below. For each question we take the retained windows, keep their number and each of their widths, and redraw their positions uniformly at random inside the recording (non-overlapping); the stored localization passes, the block ranking, the answering adapter and the per-window context stay the deployed configuration's own, so the window positions are the only difference. LEAP leads random placement by 1.41 points on TraceAV (video-clustered 95% CI $[+0.27, +2.55]$) and 4.99 on VideoOdyssey $[+2.12, +7.83]$.

C.6 The block-ranking score

Where ranking has room to act, it is measurably better than equal spacing. Block ranking only has freedom on clips long enough to tile more than three blocks, so the benchmarks separate structurally: replacing the block ranking by equally spaced blocks changes the picked blocks on 95.5% of VideoOdyssey's questions but on at most 51% elsewhere. A paired comparison on VideoOdyssey (equally spaced blocks over the same stored localization passes, identical within-block windows, answering adapter and answer scoring) costs the equally spaced variant significantly (Figure 7).

The benchmark's published per-question evidence spans say what that gain is made of. The equally spaced variant answers essentially as many window seconds per question as the deployed selection and still covers substantially less of the published evidence, a paired coverage gain of 23.07 points for ranking − equal spacing (video-clustered 95% CI $[+18.48, +27.53]$).

The contrast sharpens exactly where placement should matter most: over the questions whose annotated evidence covers under a tenth of the clip the coverage gap widens to 29.49 points, while the questions whose evidence sprawls across the clip, where almost any placement lands on it, dilute it.

What the ranking score is made of. Because σ is strictly increasing, $g_m = \max_k \sigma(\ell_{m,k}) = \sigma(\max_k \ell_{m,k})$, so blocks are ranked exactly by their maximum logit, which splits exactly into the pass's *average* logit $\bar{\ell}_m = \frac{1}{K_m}\sum_k \ell_{m,k}$, shared across the pass's windows, and the best window's *margin* over it: $\max_k \ell_{m,k} = \bar{\ell}_m + (\max_k \ell_{m,k} - \bar{\ell}_m)$. A within-pass softmax deletes the average exactly, and subtracting it deletes it while keeping the max; replacing the max by a mean keeps the average and discards the margin. The objective behind the logits matters in the same place: a selector trained with overlap-fraction BCE in place of Eq. 3 ranks blocks worse across passes. Against a twin trained on the same corpus with Eq. 3, both on a single-level grid of 75 s windows, its highest-ranked block contains VideoOdyssey's annotated evidence 15.09 points less often ($[-21.16, -9.66]$), and its top three blocks 13.87 points less often. Answered by the deployed answer pass, the two selectors' accuracies lie within a point of each other: overlap-fraction BCE − cross-entropy is -0.36 on TraceAV ($[-1.36, +0.64]$) and $+0.66$ on VideoOdyssey ($[-1.50, +2.94]$).

Deleting the average or the margin. Which ranking retains the *better* blocks is scored directly: each is read on evidence coverage over the identical stored localization passes (Table 15). Within a block the window shortlist is taken under a monotone map of the same logits and is therefore invariant to the average, so re-ranking changes which blocks are retained and nothing else. Deleting the average moves coverage by single digits, in a direction set by how it is removed. Deleting the margin, by a mean in place of the $\max$, costs tens of coverage points, and costs coverage significantly even on the questions with no partial tail block; the literal substitution $\frac{1}{K_m}\sum_k \sigma(\ell_{m,k})$ for g_m sits lowest, since with every letter's sigmoid near one it ranks a block by its *weakest* windows. What the ranking runs on is the margin. The deployed score $\max_k \sigma(\ell_{m,k})$ is its plainest form, a monotone rescaling of the strongest candidate-window logit with no normalization step of its own.

Trial-answer scores. A block can also be scored by answering the question on it and reading the chosen option's probability (Shen et al., 2025). Read on the block as the localization pass sees it, that score costs what one localization pass costs; in our pipeline it is a second pass on every block, since the window logits still come from the localization pass, taking the search from $M{+}1$ passes to $2M{+}1$. It is renormalized within its pass, so it discards the shared average as the softmax variant does. What survives is mostly about the question. A block is only ever ranked against the other blocks of its own question, so a ranking uses only the within-question part of a score's variation. Over the stored trial answers of the 30 s-window scan (§C.1) 71% of that probability's variance sits between questions, against 29% for the maximum window logit behind g_m. The two scores rank differently: on questions whose recordings are long enough to rank, they agree on the best block 30% of the time.

Put in the localization pass's place on the protocol of Table 11, with windows inside a block ranked by the base selector, the trial-answer score covers less annotated evidence than the base selector on VideoOdyssey and TraceAV and stays within a fraction of a point of its accuracy on every benchmark. The trained adapter answers more accurately than it on LVOmniBench, VideoOdyssey and MMOU.

C.7 Selection hits convert to correct answers: CG-Bench and the base-selector control

CG-Bench mini (3,000 questions over 1,118 videos of 9–105 minutes) ships *per-question human-labeled clue intervals* at second scale: an external annotation of the quantity the selector should find, on clips nearly all of which fit one whole-clip pass. Against the whole clip with the identical answer LoRA, LEAP wins the paired comparison significantly (57.50 against 49.53).

Official-protocol robustness. CG-Bench's official evaluation differs from ours in four ways: it feeds frames and subtitle text with no audio, where ours keep the native audio; its prompt carries the subtitles (with per-cue timestamps) and the sampled-frame timestamps, plus a JSON answer format with a forced-guess instruction; when no JSON parses, its letter extraction scores the gold letter anywhere in the response as correct; and both official scripts drop unanswered rows. The lenient extraction changes none of our answers, and dropping unanswered rows moves either system by less than half a point. Rerunning LEAP and the whole clip under the official prompt (for LEAP, its subtitle block cut to the retained windows takes the place of its transcript outline, without the per-frame timestamp list) and scoring with the official scripts gives LEAP 56.57 and the whole clip 50.42 on the questions each answers; the lead holds with unanswered questions counted wrong.

The human clue intervals test, descriptively, where that lead comes from. The selected windows overlap the labeled clue interval (a *hit*) on 72.3% of questions, and where they do the same answer pass is markedly more often right. Stratifying the paired comparison by this external hit criterion (Figure 9) puts most of the lead over the whole clip inside the hit stratum, and the paired flips break the same way: lopsidedly for LEAP where the windows hit, far less so where they miss.

The same ordering in the leaderboard's own aggregate (CRR). CG-Bench's aggregate counterpart to the stratification above is *CRR* $= \min(\text{long-acc}, \text{clue-acc})/\text{clue-acc}$, with long-acc the accuracy on the whole video; it requires a *clue-acc* condition: the same answer LoRA and transcript outline, with the media restricted to the human clue intervals, under the official clue prompt at the official frame count of 32, with at least two frames per clue interval. It puts clue-acc at 64.47. With long-acc taken from the paired comparison above, over all 3,000 questions, LEAP realizes $\mathrm{CRR}{=}0.892$ and the whole clip 0.768, the same ordering the stratified reading gives.

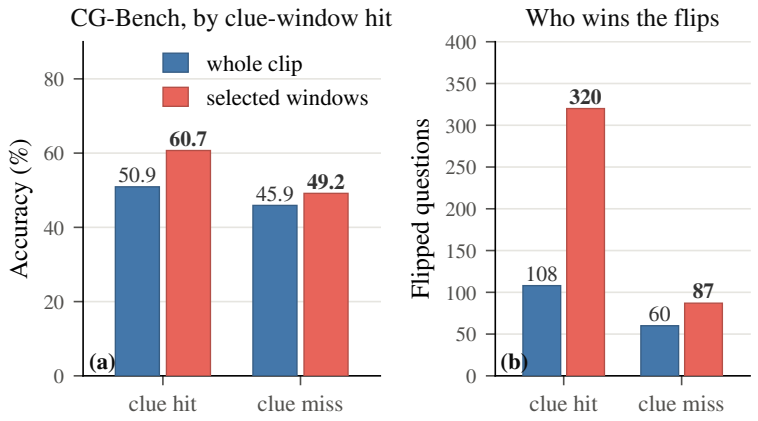

Figure 9: **Clue-grounded conversion on CG-Bench.** LEAP against the whole clip; questions split by whether the selected windows overlap the human-labeled clue interval (*hit*) or not (*miss*). *(a)* Accuracy within each stratum. *(b)* Discordant flips, where exactly one of the two is correct.

The internal companion: the same stratification inside the base-selector control. Four of the five AV benchmarks of Figure 3 ship evidence timestamps, so the same question can be asked *inside* its selector ablation, where only the selector adapter varies. Classify every question by whether a final window covers its annotated evidence under the base selector and under the localization LoRA (Figure 10). On MMOU, VideoOdyssey and OmniVideoBench the paired accuracy gain concentrates substantially on the questions where the trained selector *gains* the hit, is smaller where the hit status does not change, and *reverses* where the trained selector loses the hit, on VideoOdyssey and MMOU, most sharply on MMOU. Across benchmarks, the final-window coverage left to gain spans 8.55 to 46.42 points, and the paired accuracy gain per coverage point gained spans 0.18 to 0.44.

C.8 Audio at both stages of the pipeline

Table 16: **The audio ablation on TraceAV-Bench, at the two stages it can be read.** (a) silences the localization pass alone and scores what the selected windows retain, with no answer pass; (b) keeps the audio-on selection's windows, silences their answer pass (audio and transcript outline), and scores the answers.

(a) **What the localization pass's audio buys the selected windows.** An evidence event counts as retained when a selected window overlaps its annotated span; events are grouped by whether the annotation says hearing carries them, alone or together with seeing, or seeing alone; every question is read.

Evidence spans retained	Audio on	Muted
Carried by hearing	75.1	61.5
Carried by seeing	71.2	65.6

(b) **What the selected windows hear.** The same windows and the same answer LoRA in both conditions, cut by the benchmark's own question-modality class, its code in parentheses.

Question class	Audio on	Muted
Hearing-required (`a_`)	71.9	66.5
Cross-modal (needs both) (`av_`)	59.6	47.7
Vision-only (`v_`)	44.0	43.2
Fabricated premise (`mh_`)	76.6	74.2

Audio inside the retrieval stage. On TraceAV, silencing the *localization pass* alone returns visibly different windows: the median window-set Jaccard against audio-on is 0.43 (IQR 0.29–0.64).

The changed selection can be scored directly, without an answer pass: every TraceAV question ships an evidence trajectory whose events carry both a time range and their own modality label. Scoring each selection on whether it still contains those spans separates the two conditions cleanly (Table 16a). Muting the localization pass costs both, but it costs the hearing-borne evidence substantially more, each loss significant on its own video-clustered interval, and the audio-specific excess is significant in its own right. What the localization pass's audio buys is mostly the retrieval of the spans the annotation marks as heard, alone or together with seeing.

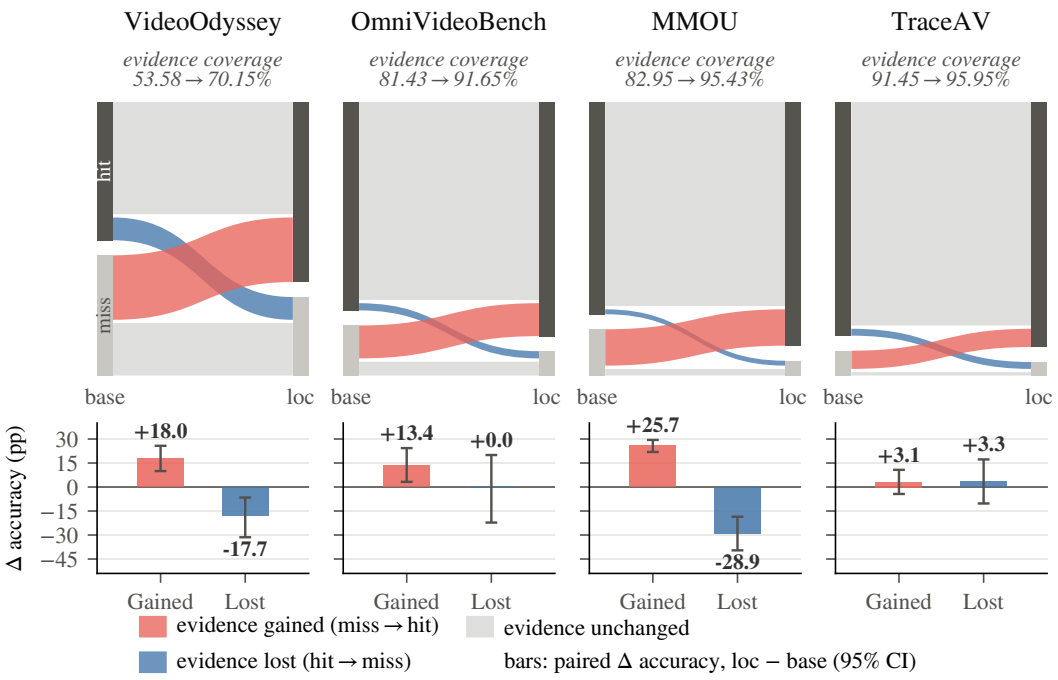

Figure 10: **Evidence-hit conversion inside the base-selector control.** *Top*: each question flows from its final-window evidence status under the base selector (left) to its status under the localization LoRA (right); ribbon thickness is the share of questions on that transition, grey the unchanged. *Bottom*: the paired accuracy change, localization minus base, on the questions of each coloured transition, with its video-clustered interval. OmniVideoBench is read on its timestamped subset, MMOU on the questions with a stored selection under both selectors; LVOmniBench ships no timestamps.

The labels those readings rest on. TraceAV is the benchmark whose questions ship those labels (answerable by hearing alone, by seeing alone, by both, or a fabricated premise the model should refuse), and its paper validates them on this backbone: its visual-only ablation reruns Qwen3-Omni-30B-A3B with audio removed at the official recipe, and the audio-centric sub-tasks fall sharply while the visual-centric ones barely move (Feng et al., 2026). We read LEAP's answer-pass silencing against that validation, over one fixed set of windows (Table 16b): the cost is largest on the cross-modal and hearing-required classes and smallest on the vision-only class.

Muting both passes, per length bucket. The muted condition, with no audio and no transcript outline, is drawn against its audio-fed twin bucket by bucket in the LVBench and Video-MME panels of Figure 13. Audio, with the outline built from its transcript, is load-bearing there, and on Video-MME its value grows with duration: muting costs several times more on the 30–60-minute bucket than on the sub-5-minute one.

D The Answering Stage and the Cross-Backbone Transfer

The no-media condition. The no-media condition of Figure 3 re-runs the answer prompts of the two selector variants without media: stem and options only, greedy, the same letter parser as every condition, the localization adapter answering as in the selector variants, and the same one-line answer instruction. The localization adapter is more accurate than this condition on every benchmark.

Localization on the smaller backbone. Figure 11 runs the selector ablation of Figure 3 on MiniCPM-o 4.5, on VideoOdyssey and MMOU: the same block grid, the same candidate-window grid and block ranking, the smaller backbone's own answer adapter in both variants, and only the localization pass's adapter swapped. Training the selector raises final-window coverage substantially on both benchmarks; block coverage moves substantially only on VideoOdyssey, where the hours-

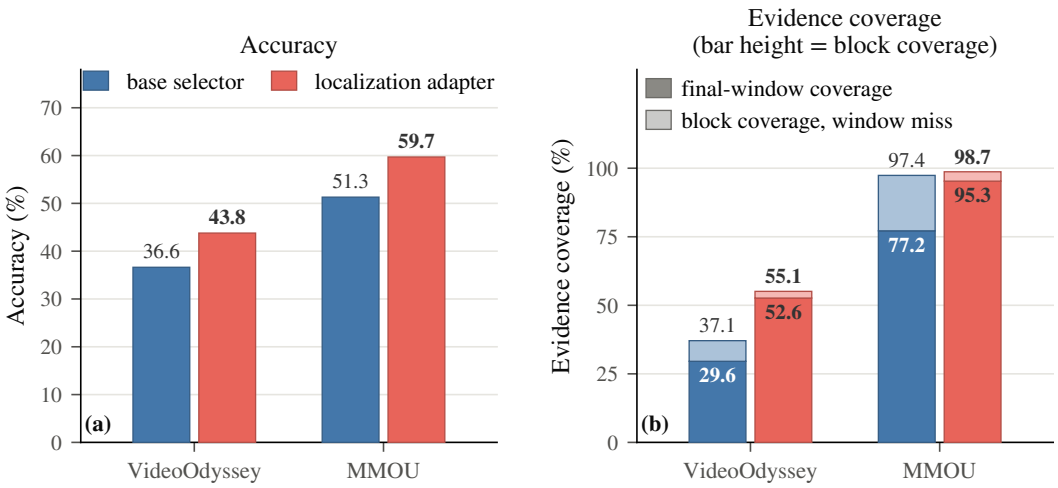

Figure 11: **The retrieval stage ablated on MiniCPM-o 4.5.** Only the localization adapter differs. *(a)* Accuracy. *(b)* Evidence coverage; full bar = block coverage, saturated part = final-window coverage.

Table 17: **The smaller backbone against its published rows and two systems of its size class.** Accuracy (%). * = published number (Feng et al., 2026; Tao et al., 2026b; Goel et al., 2026). [a] = a re-watching system on Qwen3-VL-8B with a Whisper encoder, as its paper reports it (Li et al., 2026). [b] = a native omni agent on Qwen2.5-Omni-7B, run by us under its own evaluator and step limit (Xing et al., 2026). — = no number. Columns follow each publication's protocol: TraceAV's two tables as subtask macro-averages, LVOmni micro, MMOU's test-15K split under the benchmark's own evaluator. **Bold** = best, underline = second best in each column.

System	TraceAV General	TraceAV Hallucination	LVOmni	MMOU test-15K
video-SALMONN-R^3 (8B)[a]	—	—	<u>42.9</u>	—
OmniAgent-RL-7B[b]	47.6	42.0	39.4	31.6
OmniAgent-RL-7B, as published*	<u>55.0</u>	46.1	—	—
MiniCPM-o 4.5, as published*	44.8	<u>66.5</u>	34.8	<u>46.8</u>
+ LEAP	**57.8**	**69.6**	**44.4**	**51.0**

long recordings leave the block ranking room to move, and stays near ceiling on MMOU, most of whose recordings fit one block. Paired accuracy rises significantly on both, on MMOU by more than the same swap buys the larger backbone, from a lower untrained starting point.

Against the published rows for the same backbone. Three of the smaller backbone's benchmarks publish their own MiniCPM-o 4.5 row (Table 17). MMOU withholds the gold labels of its test-15K split and scores submissions with its own evaluator, so every row of that column is the benchmark's own score. Outside this column, Figure 12 and the MMOU axis of Figure 2b, every MMOU number in this paper is on the benchmark's 5,000-question test-mini split, whose gold labels are public; the two splits share no question. We submitted the deployed MiniCPM-o 4.5 configuration on all 15,000 questions, answering each once in the original option order; questions the pipeline leaves unanswered are kept in the submission, so the denominator is the full split. The published row's protocol additionally votes over five option-order shuffles and selects among prompt variants, neither of which we use. That evaluator also returns a duration breakdown (Figure 12): LEAP is ahead in every bucket, and its accuracy on the longest videos stays near its overall one.

The agent of its size class. Our OmniAgent-RL-7B row drives the released checkpoint with its own evaluator at its defaults: a duration-adaptive cap of at most thirty-two steps, sampled decoding, and our extractor for scoring. TraceAV's appendix drives the same checkpoint with the benchmark's harness and scorer under a thirty-two-step cap (Table 17). LEAP leads both readings.

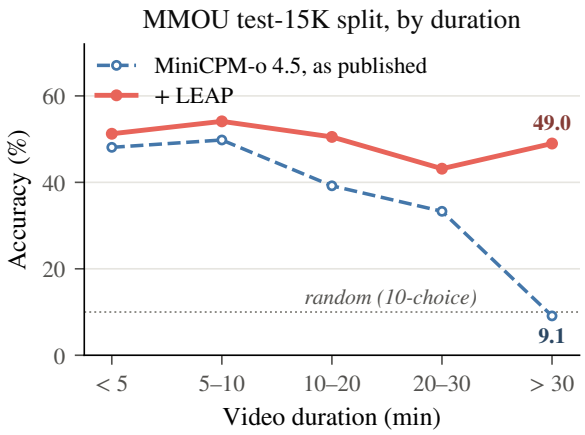

Figure 12: **MMOU by video duration: MiniCPM-o 4.5 on the test-15K split.** Scored per duration bucket by the benchmark's own evaluator. Dashed = the duration breakdown the benchmark paper reports for its own MiniCPM-o 4.5 run; solid = the same backbone carrying LEAP.

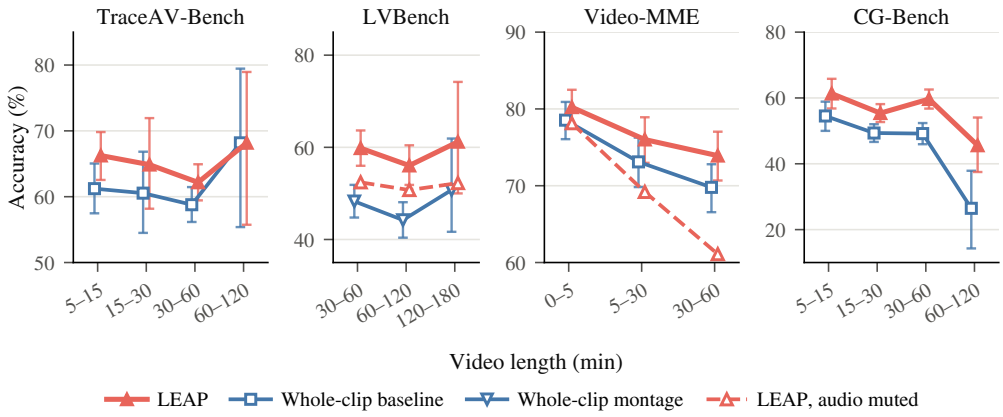

Figure 13: **Accuracy vs. video length.** Video-MME pools its two middle buckets. Muted = audio off at both passes, no transcript outline. Bars = video-clustered 95% interval, none on the muted line.

E Where the Advantage Comes From: Duration, Certificate Length, and Evidence Position

E.1 Accuracy across video length

The three video benchmarks' official protocols are visual, and LVBench's own distribution ships no audio, whereas our runs feed native audio and video, LVBench's from local copies of its source videos (video only on the questions whose copy carries no audio stream). Figure 2 reads LVBench and Video-MME with audio removed throughout; without audio every clip fits one whole-clip pass, so the baseline of Figure 2a there is the whole clip. Figure 13 reads them with native audio, its dashed line being the muted condition. Every whole-clip and montage baseline on them carries the answer LoRA. In Figure 13, each panel's baseline follows §4: the whole clip on Video-MME and CG-Bench, where $\geq$99.3% of questions fit one whole-clip pass, and the context-filled montage on LVBench, where questions exceed the whole-clip limit while the montage always fits.

On MMOU LEAP rises monotonically with duration. Up to thirty minutes LEAP shares part of that rise with the whole clip read over the *same questions*; past thirty minutes LEAP holds its level and its lead over the whole clip widens (Figure 4).

On the three video benchmarks, the per-bucket detail behind §4.6: on LVBench LEAP leads the context-filled montage in every bucket; on CG-Bench LEAP leads the whole clip in every bucket; on

Table 18: **Accuracy by certificate length on both VideoOdyssey tracks.** *Certificate length* = the benchmark's annotated length of the continuous span a viewer must watch to answer. † = the native-sampling whole clip on the same backbone, on the audio-visual track with the benchmark's official audio montage. The LEAP rows re-answer their selected windows at 14 frames per window, matching the † frame count, and without the transcript outline. **Bold** = best per column and track.

System	Certificate length (min)					Overall
	[0, 0.5)	[0.5, 3)	[3, 15)	[15, 60)	[60, ∞)	
VideoOdyssey-V (vision-only)						
Qwen3-Omni-30B base†	38.6	44.5	**44.8**	**40.4**	35.8	41.2
LEAP	**44.0**	**49.5**	41.3	**40.4**	**43.1**	**44.1**
VideoOdyssey-AV (audio-visual)						
Qwen3-Omni-30B base† (+ audio montage)	34.0	33.7	43.0	35.3	38.9	36.7
LEAP	**54.5**	**48.4**	**54.0**	**38.3**	**55.6**	**50.5**

Table 19: **The two readings of the certificate-length axis, tested on VideoOdyssey-AV.** *Certificate hit* = an answered window overlaps the annotated certificate span. *Random placement* keeps the count of LEAP's selected windows and redraws their positions over the recording. **Bold** = best in each column within a panel.

	Certificate length (min)				
	[0, 0.5)	[0.5, 3)	[3, 15)	[15, 60)	[60, ∞)
Certificate hit (%)					
LEAP	**60.7**	**69.2**	**65.8**	85.0	95.8
Random placement	9.4	20.5	43.0	**88.7**	**100.0**
Accuracy (%)					
LEAP	**55.7**	**49.6**	**53.3**	37.6	**61.1**
Random placement	47.1	39.0	47.4	**38.3**	54.2
Context-filled montage	39.3	39.3	47.8	**38.3**	47.2

Video-MME with native audio LEAP leads the whole clip in every bucket, the more so the longer the recording. The TraceAV panel of Figure 13 covers only questions inside the whole-clip limit.

E.2 Inside VideoOdyssey: certificate length

VideoOdyssey annotates every question with the length of the continuous span of the recording that certifies its answer, and it ships both a vision-only and an audio-visual track. Table 18 sorts both tracks by that length. Its rows are matched in frame count. Our variant re-answers the deployed selection with the answer LoRA at a per-window frame cap sized so that nine windows approach the native-sampling whole clip's frame count, at its pixel size, and without the transcript outline.

Both tracks are lifted most at the two ends of that axis and least in the interior, and the two ends admit two different readings, tested below. On the audio-visual track the sharpest gain sits where the certificate is shortest; one reading is evidence locality: the answer lives in a small span of a long recording, and a trained selector recovers it at the per-window context. The other is context pressure at the longest certificates: evidence that extensive is landed on by almost any placement, and what remains is the advantage of windows read at the per-window context over a baseline stretched thinnest across these recordings. The interior is where both readings predict the least.

Testing the two readings. Table 19 tests both on the deployed selection, at the deployed 32 frames per window; the LEAP row of Table 18 answers the same windows at 14 frames each, which is why the two LEAP rows differ. It sets that selection against a random placement of as many windows, drawn as in Appendix C.5 and answered without the transcript outline like every row here, and against the context-filled montage read by the same answer adapter. Evidence locality predicts that placement

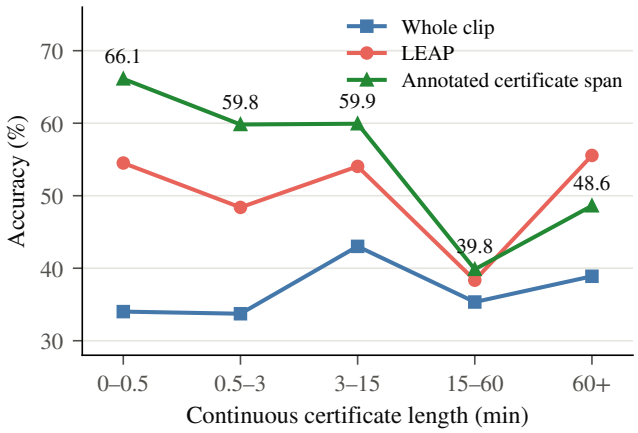

Figure 14: **The annotated certificate span on VideoOdyssey-AV, by certificate length.** *Whole clip* = the native-sampling whole clip. LEAP reads 14 frames per selected window. A moment-annotated question is absent from the certificate condition, the only one labelled with values.

matters where the certificate is short. There LEAP's windows hit the certificate far more often than random windows, and answer significantly more accurately. Past fifteen minutes random placement hits the certificate about as often. Context pressure predicts a lead that remains once placement stops mattering. At the longest certificates nearly every placement hits, the recordings are the longest of the five levels, and LEAP still leads the context-filled montage significantly.

What the annotated certificate is worth. The benchmark also asks what happens when the answer pass reads that certifying span instead of the recording, and Figure 14 runs it on this stack, sampled like the native-sampling whole clip (2 frames per second, at most 128 frames). LEAP's selected windows trail it significantly at the two shortest bins, and the gap closes as the certificate widens. Past fifteen minutes the certifying span outlasts the eleven minutes of audio an official-style montage carries, so the span arrives with the audio subsampling the native-sampling whole clip already carries.

E.3 Inside VideoOdyssey: the task types

The benchmark also labels every question with one or more of eighteen task types. Table 20 adds the LEAP row to its own per-task comparison, read against the same-stack baseline (§B.3).

LEAP's per-task accuracy is above that baseline's on every task type except Counting.

E.4 The lead over the montage across its audio coverage

The five bins of Figure 15 are cut on the context-filled montage's own audio coverage, which falls with duration. With the same answer LoRA on both, LEAP leads the montage significantly in every bin, including the one where the montage keeps almost all of the audio. With the base weights on both, LEAP still leads in every bin, significantly in the second and third.

E.5 Where the evidence sits in the clip

Figure 16 splits LEAP and the whole clip by evidence position. On MMOU LEAP climbs across the axis and leads by more once the evidence sits past the tenth minute; on CG-Bench it leads in every bucket. On VideoOdyssey LEAP does not decay across evidence positions spanning hours.

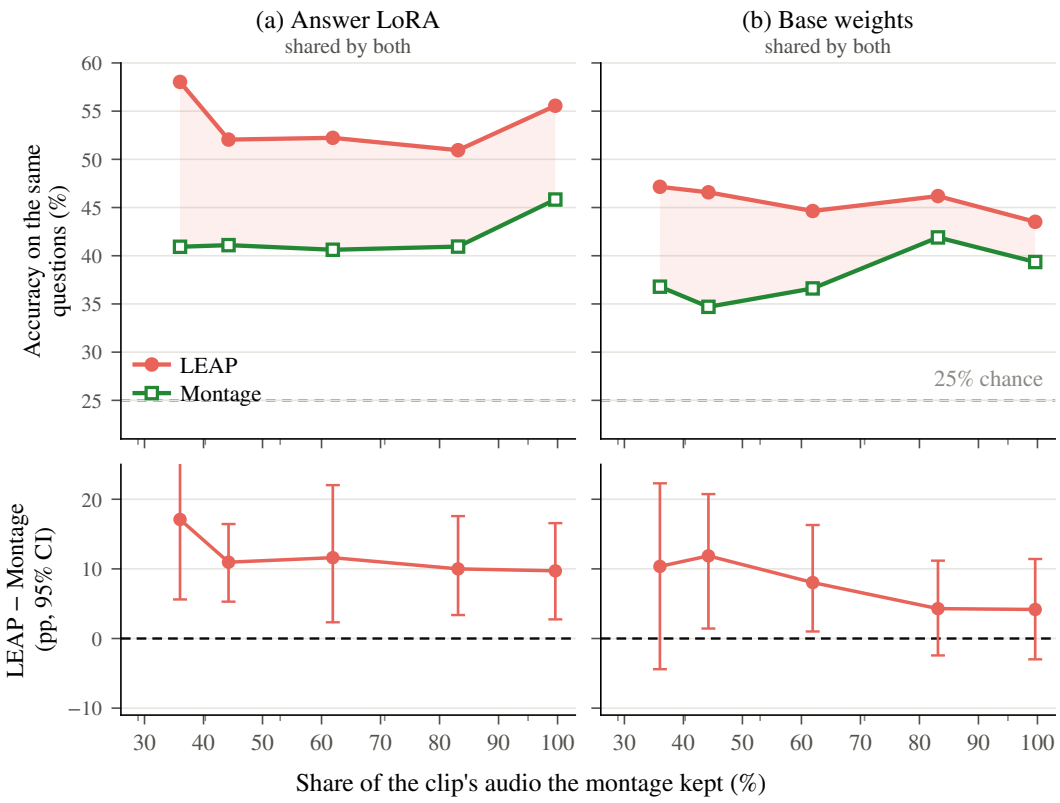

Figure 15: **Accuracy by montage audio coverage, VideoOdyssey-AV.** Questions pool into coverage quantile bins, points at bin means, edges as minor ticks. Both systems answer with the answer LoRA in *(a)* and with the base weights in *(b)*. *Bottom*: paired difference.

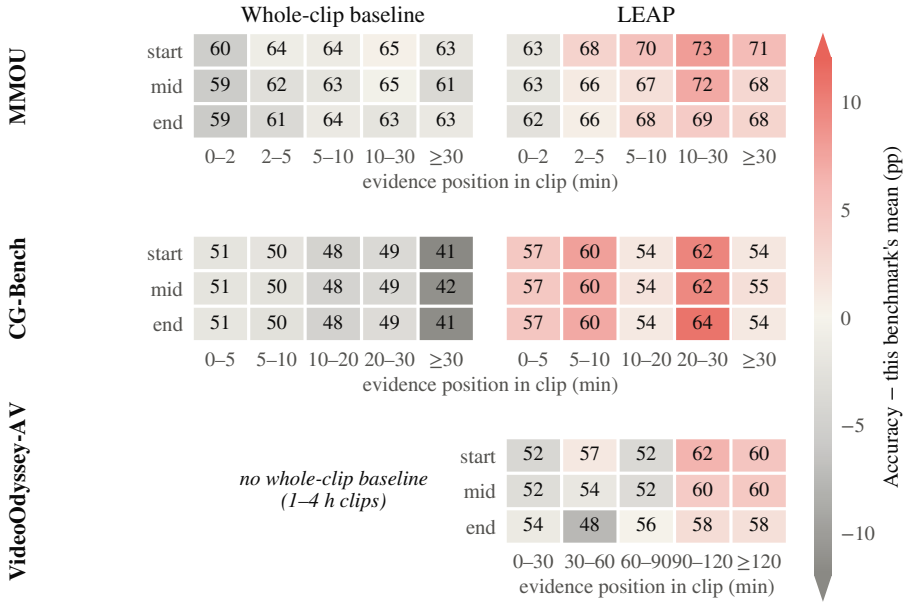

Figure 16: **Accuracy by evidence position, over the questions carrying evidence timestamps.** The rows reduce a question's annotated evidence span to one instant: its start, midpoint or end. LEAP and the whole clip carry the answer LoRA. Cells print bucket accuracy.

Table 20: **Per-task accuracy on the audio-visual track of VideoOdyssey.** Columns are the benchmark's task types; a question annotated with several is scored under each. *Count* counting, *ObRec* object, *AcRec* action, *VAR* visual attribute, *AER* acoustic event, *AAR* acoustic attribute and *OCR* character recognition, *SFR* speech fact retrieval, *Cap* captioning, *CaRea* causal, *EmRea* emotional, *InRea* intentional and *ObRea* object reasoning, *SCR* speech content reasoning, *SpRea* spatial reasoning, *Order* temporal ordering, *Sum* summarization, *TeGro* temporal grounding. ‡ = the benchmark's official input recipe re-run on our stack; rows above it are published values. **Bold** = best, underline = second best in each column, the human row excluded.

	Perception									Cognition									
System	Count	ObRec	AcRec	VAR	AER	AAR	OCR	SFR	Cap	CaRea	EmRea	InRea	ObRea	SCR	SpRea	Order	Sum	TeGro	Overall
Human baseline																			
Human	75.0	85.0	85.4	81.0	78.7	71.1	81.1	87.9	80.6	81.0	74.4	79.3	73.7	75.0	82.6	71.0	70.8	93.2	80.7
Proprietary omni-modal LLMs																			
Gemini-2.5-Pro	25.9	41.3	40.0	44.8	38.4	44.6	45.8	50.3	74.5	50.0	46.8	48.1	**42.9**	53.2	23.3	38.0	60.0	37.7	43.9
Gemini-3-Flash	30.9	44.4	45.0	36.2	32.6	46.2	45.8	50.9	66.0	56.9	51.6	48.1	**42.9**	53.2	30.0	**44.0**	50.0	32.8	44.3
Gemini-3.1-Pro	28.4	**52.4**	40.0	43.1	37.2	41.5	52.1	59.5	61.7	53.5	46.8	**65.4**	41.1	41.9	28.3	34.0	64.0	39.3	46.1
Qwen3.5-Omni-Plus	37.0	50.8	38.3	46.6	38.4	52.3	33.3	43.4	72.3	50.0	48.4	40.4	35.7	45.2	31.7	20.0	62.0	42.6	43.0
Open-source omni-modal LLMs																			
Qwen3-Omni-30B	24.7	34.9	36.7	22.4	25.6	29.2	29.2	28.9	38.3	29.3	29.0	26.9	23.2	29.0	16.7	18.0	34.0	37.7	28.7
Qwen3-Omni-30B (official recipe)‡	**38.3**	46.0	40.0	37.9	36.0	35.4	35.4	39.9	66.0	37.9	33.9	30.8	19.6	24.2	26.7	26.0	54.0	44.3	36.9
LEAP	34.6	49.2	**50.0**	**50.0**	**47.7**	**53.8**	**60.4**	**72.8**	**78.7**	**60.3**	**53.2**	48.1	39.3	**61.3**	**36.7**	40.0	**66.0**	**52.5**	**53.7**