UMMI: SYMBOLIC TIME BINDING FOR BOUNDED MULTIMODAL MEMORY WITH A FROZEN LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

A bounded memory can give a frozen language model access to multimodal sensor recordings: a small network compresses the streams into a fixed number of embeddings that are spliced into the model's input. We find that such memories support questions about what happened far better than questions about when. The limitation is not the reader alone: given the same facts as timestamped text, the same frozen Qwen3-Omni model places events in the correct interval with 77–78% accuracy without any training. We introduce symbolic time binding in UMMI, a learned memory interface. Memory embeddings are grouped by recording interval, and each group is prefixed, in ordinary text, with its interval index and step range. The backbone is not updated and the continuous memory is not enlarged; the prompt grows by sixteen short phrases. On a synthetic dense-event benchmark, interval localization rises from 60.7% to 98.0% over three training seeds, whereas learned interval embeddings added to the same tokens help much less. On question subsets from held-out OPPORTUNITY subjects, overall accuracy rises from 41.6% to 53.8% for the same shared-memory configuration, with gains that vary across subjects. A flat tokenizer given the same labels does not benefit. Time binding also fails to help in the tested Aria Everyday Activities configuration, where several transcript words share each interval. Explicit textual time labels are thus a simple and effective interface for reading time from a bounded memory, while preserving several items within one interval remains open.

1 INTRODUCTION

Sensor foundation models are judged on recognition. EEG-FM-Bench (Xiong et al., 2026) spans fourteen datasets and every task in it is a classification; the multimodal long-term-memory benchmarks that exist (Ren et al., 2026; Wu et al., 2025; Maharana et al., 2024; Bei et al., 2026) are conversations of text and images; egocentric benchmarks (Grauman et al., 2022; Yang et al., 2025) ask about the past through frames. We study whether a frozen language model can answer questions about past events in heterogeneous sensor recordings using only a bounded memory. The obvious interface for that setting is a bounded memory: a fixed number of embeddings written from the streams and spliced into a frozen language model's input, the way sensor encoders are spliced today (Li et al., 2025a; Zhang et al., 2025b; Sivaroopan et al., 2026) and the way bounded memories are built for single streams (Ryoo et al., 2023; Liu & Liu, 2026; Zou & Donz, 2026). Two alternatives fail measurably: an unbounded context loses 20 points when the same evidence is padded to 40k tokens (LongMemEval, Appendix B), and rank-32 LoRA on the backbone leaves the chain task at 5% (§7). We built such a memory, on a frozen Qwen3-Omni-30B, and it worked on the tasks we designed it for. Then we asked it *when* something happened, and it could not say.

We investigate why these memories support some event queries but struggle with interval localization. On Aria Everyday Activities, a store that holds the speech transcript alongside video and inertial streams answers which word was said first at 88% and in which of eight intervals a word was said at 19%, five points above its majority rate of 14, and ten epochs do not move it. On a synthetic set of dense events, built so that the answer is a known function of what the memory holds, the store names the first event at 91% and places it in one of sixteen intervals at 9%. And the same frozen model, given the same facts as a timestamped transcript, places them at 77–78% with no training at

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.

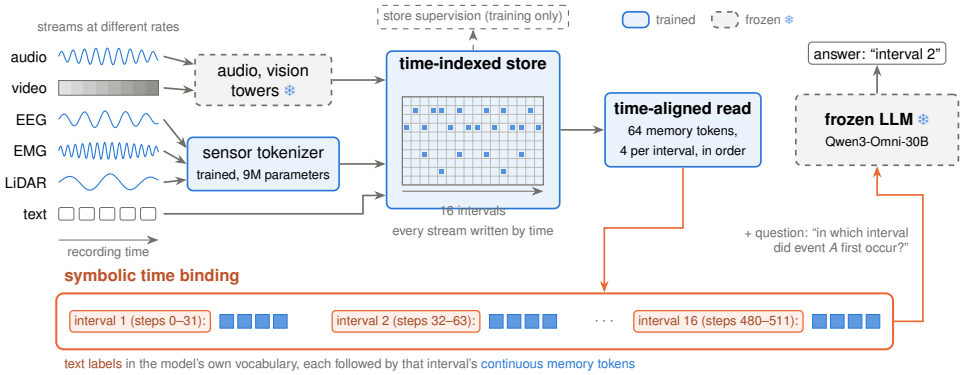

Figure 1: UMMI. Streams at different rates are encoded by frozen towers and a trained sensor tokenizer and written by time into one store. The time-aligned read returns 64 memory tokens in recording order, and symbolic time binding places each interval's tokens after a text label giving its ordinal and step range. The language model stays frozen.

all (§4). The pattern recurs, to varying degree, across the stores we tried. The memory delivers time as the position of an embedding in the prompt. The frozen model reads sequence position natively, but under our training it does not reliably map that position onto the absolute time of a recording.

The method follows from the diagnosis (§5). The memory is read interval by interval, so each interval's tokens can be bound to a symbol the model already reads: the interval's ordinal and step range, written in the model's own vocabulary immediately before the tokens. The store architecture, the read and the objective stay as they were; the memory is trained with the labels present and the prompt gains sixteen short phrases. On the dense-event set the interval question rises from 61 to 98% and all questions from 68 to 89% (§6). On held-out OPPORTUNITY subjects the label-free shared store rises from 41.6 to 53.8% on identical question subsets, unevenly across subjects, and on PAMAP2 every time-indexed memory gains 6–9 points over all questions of three held-out subject pairs (label-free store 68.7 $\rightarrow$ 74.9%; §8). Two controls delimit where the symbols help. A flat tokenizer, whose 64 tokens are a learned compression of the whole recording, does not benefit from the same symbols. And on Aria, where about seven transcript words share each interval, the symbols do not help; mixing several words within one interval is a plausible cause that we have not isolated.

The store beneath the symbols is a single time-indexed store shared by every stream, with a time-aligned read, time-sliced encoding wherever a frozen tower is not temporal, and supervision on what each interval holds, with or without labels. §7 measures each part on a six-modality chain benchmark; distractor-cost measurements on three published memory benchmarks are in Appendix B.

Contributions.

1. We characterize interval-localization failures in bounded memories read by a frozen language model, using a tower-free dense-event benchmark and timestamped-text controls (§4).
2. We introduce symbolic time binding and evaluate textual interval labels against positional memory order and learned interval embeddings; interval localization on the dense-event set rises from 60.7 to 98.0% (§5–§6).
3. We examine the method on real sensor recordings (PAMAP2 68.7 $\rightarrow$ 74.9%, OPPORTUNITY 41.6 $\rightarrow$ 53.8%, each on matched questions) and report its failure in the tested Aria configuration, alongside a component study of the underlying multimodal memory (§7–§8).

2 Related Work

Sensor-to-language models do perception, not memory. SensorLM (Zhang et al., 2025b) and SensorLLM (Li et al., 2025a) for wearable motion, OpenTSLM (Langer et al., 2026) for medical time series, NeuroLM (Jiang et al., 2025), BELT (Zhou et al., 2024) and TFM-Tokenizer (Pradeepkumar et al., 2026) for EEG, and more recently SLIP's language-informed sensor pretraining (Chen

et al., 2026) and STELLA's fixed-budget inertial tokenizer (Sivaroopan et al., 2026) all map a signal into a language model's embedding space and tune the pair; their benchmarks are activity recognition, sleep staging and captioning, judgements about *a* window. None asks the model to keep a stream over a long horizon and answer later about a moment it no longer sees, which is the setting of §3.

Bounded memories and learned-query readers over a frozen LM. The free-query read that loses on the chain task is the latent-bottleneck reader of Perceiver (Jaegle et al., 2021) and BLIP-2's Q-Former (Li et al., 2023); Memorizing Transformers (Wu et al., 2022), RMT (Bulatov et al., 2022), Infini-attention (Munkhdalai et al., 2024), Titans (Behrouz et al., 2025) and the streaming video memories (Song et al., 2024; He et al., 2024; Zhang et al., 2025a) keep a bounded store over one stream at one rate, and Token Turing Machines (Ryoo et al., 2023) address it by learned position embeddings. The multimodal agent memories of the last year—WorldMM's episodic, semantic and visual stores for long video (Yeo et al., 2026), the verbatim-to-gist pyramid of Lian et al. (2026), and the systems benchmarked by MemLens (Ren et al., 2026) and Mem-Gallery (Bei et al., 2026)—keep text, frames or retrieved clips and are read by retrieval; MemLens's own finding, that memory agents lose visual fidelity under storage-time compression, is the failure this paper measures for continuous slots and time. Several streams at several rates in one bounded store outside the window, read by a frozen model, is the case here.

Time as text. TimeChat, VTimeLLM and LITA (Ren et al., 2024; Huang et al., 2024a;b) put timestamp tokens beside frame features and ChronusOmni (Chen et al., 2025b) beside audio and video, all inside the context window with the model fine-tuned to use them; Qwen-Omni's TMRoPE aligns modalities by time positionally. Symbolic time binding puts the timestamp beside a *memory* token on a *frozen* model, where positional time is what fails (§4). Video-centred agent memories, context compression into learned tokens, importance-ranked retention and EEG foundation models are discussed in Appendix F.

3 SETTING

Backbone and interface. The backbone is Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), frozen throughout. Its *thinker* is the language model with its audio and visual towers, used as shipped; its *talker* and *code2wav* vocoder turn the thinker's output into speech and are loaded, not removed (Appendix C). Sensor streams enter through a learned path of 9.0M parameters: a per-channel convolutional stem, a transformer over (channel, timestep), a bounded *store* of 256 slots, and a *read* that produces 64 embeddings spliced into placeholder positions of the thinker's input. Audio and image features are either spliced the same way onto their own placeholders, sized from what each tower actually emits, or—in the *memory-only* condition used for every result from §4 on—written into the store by time, so that nothing from the towers enters the prompt. The store is computed once per recording and is independent of any question; only the read sees the question. Two stores are used below: the *shared* store, one time grid that every stream writes into, and the *question-blind* store, the same slots and read with the question vector zeroed and one grid per sensor path. The two are compared as configurations; they differ in sharing and, below, in supervision, and we do not separate the two factors.

The chain task. A recording is 512 steps of EEG, EMG and LiDAR, 25.6 s of audio, and one rendered strip whose sixteen cells are consecutive time buckets. Six cues sound during it. Each cue's *time* is audible only; the *colour* lit at that moment is visible only; whether the subject was engaged, whether a muscle fired, and whether an object was near are in EEG, EMG and LiDAR respectively (signal parameters in Appendix C). A rule maps (colour, engaged, muscle, near) to one of sixteen outcomes, and the question names a cue by ordinal, so the other five are interference of the same kind as the target. Seven question families—four single facts, depth-2, depth-4 and open—share every recording (Table 9, appendix); retrieval is identical across them, so what changes between families is attributable to the question. The set has 3,000 recordings and 24,000 questions, split $80/20$ by recording, never by question; every chain number is the last training epoch on the held-out recordings, with no epoch or seed selection on them. The chain benchmark was the store's *development* set (the choices of §7 were made on its held-out recordings), so we report it as a component study, not a test; the binding was designed on the diagnosis of §4 and evaluated once,

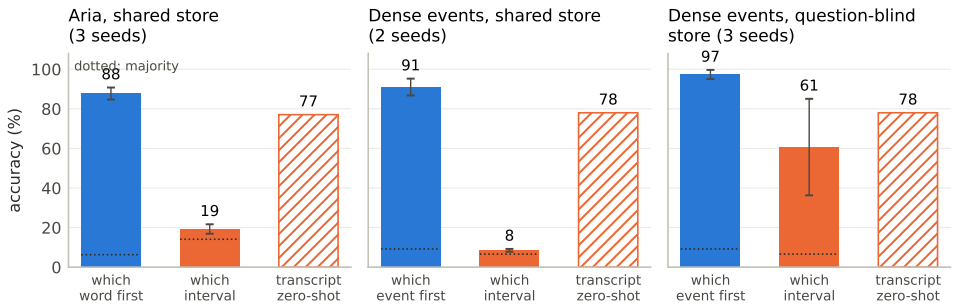

Figure 2: Interval localization lags event recognition. Unbound memories (solid) name the first item well above majority (dotted) but place an item in its interval near majority; the same frozen model reading the same facts as timestamped text (hatched) places them zero-shot. Error bars: s.d. across seeds.

at fixed settings, on the dense-event set and the real corpora, whose held-out subjects chose nothing (the symbol count, one per interval, was fixed before the sweep of §8). Closed families are scored by the first generated token; where a backbone does not answer in one token we rank the family's closed answer set by length-normalised likelihood (*likelihood-ranked*).

Audits. Colour selects the outcome table, so no depth question is answerable without vision. For the fusion families a single-view oracle that knows one modality perfectly scores below majority, so both streams are required. A probe trained on the raw window with the cue position given reads engaged at 100% and muscle at 99.7%, against 47.7% and 64.3% when the position is shuffled: the facts are present and localised. Scores are reported with the majority class beside them, and with balanced accuracy wherever a family is imbalanced.

The dense-event diagnostic set. The chain task places one cue per interval, so a memory that averages an interval loses nothing. To test the memory where averaging must fail we built a second synthetic set with no towers at all: 512 steps of seven sensor channels carrying short events of twelve types (each a fixed eight-step multichannel waveform on coloured noise, with amplitude and duration jitter), one to four events per 32-step interval, about forty per recording, from a per-recording subset of six types. Six closed families ask in which of the sixteen intervals the first event of a type occurred (`when`, 16-way, majority 7%), which event came first, which followed the first of a type, how many times a type occurred, whether A preceded B, and which event was closest to a step (`state-at`); 3,000 recordings, 24,000 questions, eight synthetic subjects, one held out. The supervision head sees three bits per interval (more than one event; type change from the previous interval; energy above median) and never an event's identity or step. Every answer is a function of the event list, and a transcript of that list gives a perception-then-text reference (§6).

The three real corpora (PAMAP2, OPPORTUNITY, Aria Everyday Activities) are described with their results in §8.

4 Diagnosing interval localization in bounded memory

Every result in this section uses the store of §7 trained to convergence, with the 64 memory tokens as the model's only access to the recording (no tower output in the prompt), and reads accuracy *per question family* rather than in total. The families split into those that need an item recognised and those that need it placed in time, and the two groups behave differently (Fig. 2).

Aria: which word is read, when it was said is not. On Aria Everyday Activities the transcript words are written into the store at their time stamps alongside video and inertial streams (§8). Over three training seeds on the held-out location the memory answers which of the words was said first at $87.7 \pm 3.0\%$ and which was said after a named word at 69%, but in which of eight intervals a word was said at $19.3 \pm 2.4\%$ against a majority of 14.1, and a ten-epoch run (not shown) does not

move it (17%). These runs use an earlier build of the corpus (253 held-out questions); on the full build (1,378 held-out questions, one seed) the pattern is the same, 61% against 14% (majorities 1 and 11; Table 5). The store holds the words, but the model does not recover where in the recording they occur.

The dense-event set: the failure is reproducible, and the answer is in the memory. On the diagnostic set the same store names the first event at $91 \pm 4\%$ and places the first event of a type in its interval at $8.5 \pm 0.7\%$, chance being 6.6%; a bucketed mean-pool of the tokenizer output, no learned slots, reads it at 33–64%, and a question-blind learned store with the same supervision at 33–79% across seeds (Table 1). That the memory carries the time is shown by the mean-pool: its tokens are nothing but per-interval averages in interval order, so the item and its interval are both in the prompt, and the model still misplaces one event in two.

The frozen model reads time when it is written as a symbol. The last control needs no training. Given the Aria transcript as text with each phrase's interval and step range written before it, the same frozen backbone answers the interval question at 77.0% and which-came-after at 78.1% zero-shot (Table 5); given the ground-truth event list of the diagnostic set as such a transcript, it places an event in its interval at 78% and answers 70% of all questions. The frozen model can locate a moment when the moment is a token it reads; it cannot when the moment is the position of an embedding. The perception-then-text rows of Table 3 show the same effect from the other side: a ground-truth activity transcript with no interval marked is read at 53%, and telling the model which interval to look in raises it to 60%, the two transcripts differing by one phrase.

What changing the store did not fix. These controls suggest that the representation of time at the memory–model interface contributes to localization errors. Learned slots, label-free supervision, per-stream stores and an event-indexed store (Appendix E) all sat within seed spread of the mean-pool on the real corpora; none of them changes how time reaches the model, which is as the index of a token. Freezing the backbone still allows the input interface to learn through it, so our claim is empirical: under the tested training setup, explicit textual interval labels are more effective than relying on memory-token order or adding learned interval embeddings (Table 11).

5 SYMBOLIC TIME BINDING

The mechanism. The store holds 256 slots in recording order, sixteen per interval of 32 steps. The time-aligned read produces 64 tokens: token i ($i = 1..64$) is the mean of slots $4i - 3$ to $4i$ (a quarter of an interval) plus a learned query vector $\mathbf{q}_i$, to which the shared store adds a linear projection of the question embedding (zero in the question-blind store), followed by layer normalisation and a linear map into the backbone's embedding space. Symbolic time binding groups the 64 tokens into N consecutive groups ($N = 16$ here, four tokens per interval) and writes, immediately before each group and in the backbone's own tokens, the group's ordinal and step range:

$$\texttt{interval 3 (steps 64-95):} \quad \langle m_9\rangle\langle m_{10}\rangle\langle m_{11}\rangle\langle m_{12}\rangle$$

The store architecture, readout and training objective follow §7. The memory is trained with the interval labels present. Its continuous budget remains 64 embeddings, while the prompt gains sixteen text phrases. Each memory token now sits beside a discrete symbol the model has read since pretraining, so a question that names an interval ("in which interval", "at step 287") can be matched to that symbol by the frozen attention, and an answer that is an interval can be copied from it. The memory stays bounded and continuous; only its address becomes a word.

Why this is a binding and not a prompt format. The symbols carry no information about the recording; they are the same sixteen phrases in every prompt. What they change is what the model can do with the embedding beside them, and the controls of §6 make that precise. Bound to tokens that are time-local and supervised, the symbols take the interval question from 61 to 98%. Bound to a flat tokenizer, whose 64 tokens are a learned compression of the whole recording, the same symbols leave it at 4% and lower the other families. The results support pairing textual interval labels with memory tokens organized by time. Adding learned interval embeddings to the memory tokens yields smaller and more variable gains than textual labels on the diagnostic set (when $60.7 \rightarrow 71.3\%$

	when	first	after	count	before/after	state-at	all
	accuracy (%), held-out subject, last of three epochs						
majority	6.6	9.2	9.3	30.5	50.4	8.9	–
perception→text: ground-truth event transcript, zero-shot	78.0	88.0	26.0	70.0	88.0	66.0	70.0
flat tokenizer + time symbols	4.0	18.0	30.0	28.0	54.0	9.0	19.5
bucketed mean-pool, 2 seeds	48.5±21.9	96.0±0.0	41.0±1.4	53.0±9.9	87.0±4.2	49.5±6.4	59.1±3.4
+ time symbols, 2 seeds	80.0±4.2	99.0±1.4	44.0±8.5	56.0±2.8	88.0±2.8	68.0±4.2	72.9±2.3
question-blind store, per-interval supervision, 3 seeds	60.7±24.4	97.3±2.3	72.7±28.4	64.0±14.0	88.7±4.6	48.3±0.6	67.6±12.0
+ time symbols, 3 seeds	**98.0±2.0**	**99.3±1.2**	**92.7±3.1**	**72.0±10.4**	**94.7±1.2**	**79.0±1.0**	**89.1±0.6**
shared store, label-free supervision, 2 seeds	8.5±0.7	91.0±4.2	38.0±2.8	31.0±9.9	62.0±11.3	44.0±1.4	40.9±2.3
+ time symbols, 3 seeds	46.0±44.5	72.7±22.1	46.0±22.5	42.7±10.3	69.3±23.2	48.3±22.4	52.4±25.6
Qwen3.8-27B backbone, question-blind store, 1 seed	8.0	94.0	36.0	28.0	54.0	42.0	39.0
+ time symbols, 1 seed	**94.0**	**96.0**	**72.0**	**60.0**	**94.0**	**76.0**	**82.8**

Table 1: Dense-event diagnostic set, accuracy (%) on the held-out subject. Each "+ time symbols" row differs from the row above only in the sixteen text labels, present in training and evaluation; the continuous memory stays at 64 embeddings. ± is the sample s.d. across seeds. The last two rows repeat the question-blind pair on a second frozen backbone; protocol in Appendix A.

against 98.0%; Table 11, appendix). This control is inspired by learned memory-row positions (Ryoo et al., 2023).

Relation to prior work. Textual timestamps beside per-interval features are the standard recipe of video temporal-grounding models—TimeChat, VTimeLLM and LITA (Ren et al., 2024; Huang et al., 2024a;b) interleave time tokens with frame features and ChronusOmni (Chen et al., 2025b) does so for audio and video—but all of them place the recording's own tokens in the context window and fine-tune the model to use the timestamps; Token Turing Machines add learned position embeddings to memory rows (Ryoo et al., 2023). We study textual time labels at the interface between a bounded memory and a frozen language model. Our experiments compare these labels with memory-token order and learned interval embeddings, and examine where the method succeeds and fails.

6 Results on the diagnostic set

Down the `when` column of Table 1, the symbols take the interval question from 61 ± 24 to $98 \pm 2\%$ on the question-blind store and from 49 to 80 on the mean-pool; without them no memory places an event reliably. Across a row, the families that need a moment located (`when`, `after`, `state-at`) move by 20–37 points, while `first`, a relative-order task already near ceiling, changes little. All-question accuracy on the question-blind store goes from 67.6 ± 12.0 to $89.1 \pm 0.6\%$, and the seed spread collapses. For reference, the frozen model answers 70% of questions when given the ground-truth event list as timestamped text. This is a zero-shot condition, whereas the memory interface is trained for three epochs, so the comparison does not isolate the effect of representing the recording as text or memory.

The flat tokenizer with symbols is at 19.5%: its 64 tokens are a learned compression of the recording with no time-locality, and the symbols lower the families it could otherwise answer. The shared store, which sums every stream into one grid, is unstable across seeds on this single-stream set with or without symbols ($81.8/34.5/41.0\%$ over three seeds with them). With one stream there is nothing to share, and we use the question-blind store here. The two stores also differ in supervision, so we do not attribute the gap to sharing alone. The real corpora below use both, and the symbols help each.

A second backbone. We repeat the question-blind pair on a frozen Qwen3.8-27B (Qwen Team, 2026), a dense vision–language model with no audio tower (Appendix C), with the same data, seed and held-out questions (Table 1, last rows). Without symbols the diagnosis reappears: the memory names the first event at 94% and places an event in its interval at 8%, the majority rate. With the symbols the interval question reaches 94% and all-question accuracy rises from 39.0 to 82.8%. This backbone fits one example per step on one GPU, so the learning rate is halved; at the original rate the unbound run is unchanged (38.8%) and the bound run diverges (Appendix A).

7 THE STORE BENEATH THE SYMBOLS: THE CHAIN BENCHMARK

This section measures the store on the six-modality chain task of §3, where evidence can be placed exactly, and reports what each part is worth; Appendix C has the full account, the component ladder (Fig. 5) and the leave-one-out ablation (Table 8).

The facts are recoverable; the model does not receive them. A probe given the cue position reads the sensor facts from the raw window at ceiling. The flat tokenizer delivers them to the frozen model at their majority class and depth-4 at 8%, and neither 512 uncompressed tokens, rank-32 LoRA on the backbone nor conditioning the read on the question changes that (Table 8). Told the four facts in words the frozen model applies the rule at 64%, so the reasoning is present and the bottleneck is delivery.

Address the store by time. A store-supervision head predicting, for each of sixteen intervals, whether a cue occurred and what its fields were, was trained with the language model removed. Read by free learned queries, every field sat on its floor for nine epochs; read interval by interval, the same store reached the probe's ceiling. Trained jointly the supervision is dose-dependent (Fig. 6), annealing it to zero loses nothing, and predicting masked intervals from the others, with no labels, keeps depth-4 at 90%. The supervision is scaffolding for the write and can be removed from the objective.

Encode per interval wherever the tower is not temporal. With the whole strip through the frozen visual tower colour plateaued at 44%: recolouring one cell at a time shows that the tower merges neighbouring time-ordered cells into shared output tokens, while the audio tower is exact. Slicing the strip along time before the tower and scattering the slices onto the store's grid takes depth-4 from 54 to 96% (five seeds). A temporal-fidelity score for every encoder, and its relation to recall, are in Appendix D.

What is load-bearing. In the leave-one-out ablation (Table 8) the time-aligned read, the store supervision and the shared store are each worth more than fifty points of depth-4; sliced visual tokens help only in the *same* store as the sensor tokens. A bucketed mean-pool with no learned slots reaches $76 \pm 22\%$ against the store's 96 ± 4: indexing by time does most of the work and the slots make it reliable. With the memory as the model's only access to the recording depth-4 remains at 88–98%.

Rules never seen, and what the control says. Trained with a different rule table per recording, so the only way to answer is to read the rule in the prompt and apply it to the recalled fields, the memory reaches 96–100% depth-4 on a rule it has never seen. A control narrows the claim. The same fields written in words, plus a 64-embedding soft prefix (Lester et al., 2021) trained under the same regime, also apply never-seen rules at 100%. The soft-prefix control also succeeds, so this result does not establish a reasoning capability unique to the memory; the store's contribution is delivering the fields. The answer is also spoken, with no retraining (Appendix C).

8 REAL SENSOR STREAMS: PAMAP2 AND OPPORTUNITY

The chain task is synthetic by design, so that evidence can be placed. To check that the same memory reads a real recording, we take PAMAP2 (Reiss & Stricker, 2012): nine subjects wearing three inertial units (hand, chest, ankle) and a heart-rate monitor through a fixed protocol of twelve activities, about an hour each. We keep seven channels (heart rate, and acceleration and gyroscope magnitudes per unit) at 3.3 Hz, cut 2.5-minute recordings of 512 steps starting every 128 steps (consecutive recordings overlap; subjects, not recordings, are held out), and ask five closed families about each: what the subject was doing at a step (13-way, including the transient class), which activity came first, what followed a named activity, how many distinct activities occurred, and whether A happened before or after B; 518 recordings, 3,749 questions. The store-supervision head sees only three bits per bucket (locomotion or not, a change of activity, heart rate above the subject's median). Held-out *subjects*, never recordings, form the test split, so no test signal overlaps a training one; a probe confirms the activity is readable from the raw window (85% balanced against 12% majority).

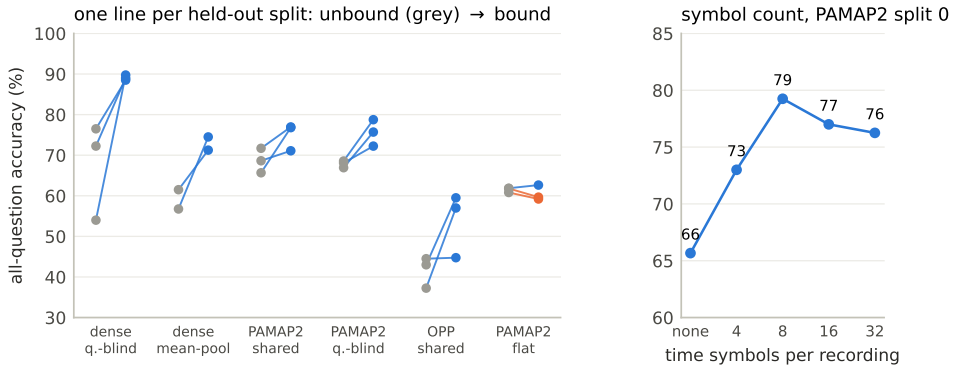

Figure 3: Left: all-question accuracy without (grey) and with (coloured) time symbols, one line per held-out split or seed, both ends scored on the same questions. Time-indexed stores rise; the flat tokenizer does not. Right: accuracy against the number of symbols on one PAMAP2 split.

	state (13-way)	first	after	count	before/after	all
	balanced accuracy (%)					acc. (%)
chance (1/classes present)	9.6	11.1	12.5	33.3	50.0	–
text: classifier transcript (2 splits)	35.4±6.5	62.4±1.7	39.6±3.2	51.7±3.0	62.1±4.3	40.1±3.2
text: ground-truth transcript + interval	52.7±2.7	88.3±1.3	58.6±2.3	77.2±3.1	66.8±2.0	60.5±0.6
flat tokenizer	54.2±5.7	55.7±5.1	92.5±1.1	49.7±0.6	89.7±3.8	61.5±0.6
mean-pool, no learned store	61.5±6.5	66.2±2.2	90.8±6.7	60.6±7.2	90.8±5.0	68.1±5.4
shared store, label-free	62.5±6.5	65.5±6.3	92.7±5.8	56.6±0.6	88.9±1.1	68.7±3.0
shared store, labelled	62.5±7.4	67.0±6.4	88.5±9.1	60.3±1.3	86.3±1.7	68.3±3.7
flat tokenizer + time symbols	50.8±1.0	54.9±6.6	89.5±2.1	53.5±4.8	82.8±5.0	60.5±1.9
mean-pool + time symbols	72.4±4.7	69.8±9.2	89.8±3.1	**63.9±6.5**	92.0±3.0	76.2±4.5
shared store, label-free + time symbols	71.1±5.9	69.0±7.1	**93.3±2.8**	56.7±1.7	92.8±2.8	74.9±3.3
shared store, labelled + time symbols	**74.5±1.9**	**70.2±6.4**	88.6±5.2	61.2±5.3	**94.5±1.6**	**77.1±2.8**

Table 2: PAMAP2: balanced accuracy per family and plain accuracy over all questions of the held-out subjects (three splits of two subjects; ± is the sample s.d. across splits). The two text rows replace the memory with a 16-interval activity transcript (perception→text). Rows below the rule add the sixteen time symbols and are scored on the same questions as their unbound rows. All rows, including the question-blind and per-stream stores, are in Table 3.

The upper half of Table 2 sets the memory of §7 against the systems it has to beat. The practitioner's pipeline—a per-window activity classifier (79% balanced on held-out subjects) whose labels are written as a sixteen-interval transcript into the frozen backbone's prompt—answers 40.1% of the questions; given the *ground-truth* transcript, 53.1% (Table 3); told in addition which interval the queried step falls in, 60.5%. One phrase naming the interval is worth seven points, consistent with the diagnosis of §4 on real data. The transcript has one label per interval because that is the corpus's label resolution; the diagnostic set gives the uncompressed forty-line event list as a reference (§6). Every memory indexed by time—mean-pool, per-stream stores, the shared store under any supervision—sits at 65–69% on the same three pairs (Table 3), seven points above the flat tokenizer and eight above the best transcript. The margin tracks indexing by time more than the slots or the labels.

Below the rule the memories are trained and read with the time symbols (Fig. 3), and every pair is scored on identical questions. All four time-indexed memories (the question-blind store is in Table 3) gain on all three held-out pairs: the label-free shared store rises from 68.7 to 74.9%, and the labelled store, the question-blind store and the mean-pool by about eight points each. The gain is concentrated where a moment must be located: state-at-a-step rises by 9–12 points in every configuration, while *count*, which needs no moment located, moves by at most three. The mean-pool gains as much as the learned stores, so on this corpus the symbols do most of the work. The flat tokenizer with the same symbols does not move. The number of symbols is not delicate: four to thirty-two give 73–79% on the first split, and the finest is not the best. On the chain benchmark of

§7 the memory-only model bound to sixteen symbols reaches 100% on colour, depth-2 and depth-4 (Table 8).

Where the binding stops: Aria Everyday Activities. Aria Everyday Activities (Lv et al., 2024) is the cross-modal case: egocentric video (one frame per interval through the frozen visual tower), two inertial units and gaze through the sensor tokenizer, and the speech transcript as a sparse text stream stamped by time, all in one store; questions are built from the transcript (in which of eight intervals a word was said, which came first or next, how many turns, before/after) and from room changes; 926 clips of 77 s, $6{,}151$ questions, held out by location. Table 5 (Appendix A) reports a negative result. With and without the symbols scored on the same $1{,}378$ held-out questions, eight symbols move the interval question only from 14 to 18% on the shared store (majority 11%), while the frozen model given the timestamped transcript answers at 77% zero-shot. The tested Aria configuration thus shows no clear localization benefit from time labels. A clip carries 54 words, about seven per interval, and the store sums them into one interval's slots; mixing several words within an interval is a possible source of information loss, but we have not isolated this effect. Item-level memory is one direction for testing the explanation.

A second corpus: OPPORTUNITY. OPPORTUNITY (Roggen et al., 2010) is the harder case: four subjects, five locomotion states from back and arm inertial units, activities that alternate every few seconds, and a raw-window probe that reads the state only weakly (43% balanced vs. 27% shuffled). Built the same way (568 recordings, 4,618 questions, held-out subjects), it gives Table 4 (Appendix A). Without symbols the best memory (52.9%) beats the classifier transcript and the flat tokenizer and only ties the ground-truth transcript given its interval (53%). With the symbols, scored on the same questions, the label-free shared store rises from 41.6 to 53.8%, by $+19.8$, $+16.5$ and $+0.3$ on the three subjects, and the mean-pool from 52.0 to 63.6%. The gain is real but uneven across subjects. The flat tokenizer with symbols does not benefit here either.

9 Conclusion and limitations

Textual interval labels improve temporal localization when a frozen language model reads the tested memories organized by time. The labels preserve the continuous-memory size but add text to the prompt. The evidence is limited to two backbones, the second on the diagnostic set with a single seed, and to a small number of seeds and subject splits, with uneven gains across subjects on both inertial corpora. The shared store is unstable across seeds on the single-stream set. The tested Aria configuration remains difficult, motivating further study of how memories preserve multiple events within an interval.

Reproducibility statement

Every number is read from a result file written by the run that produced it, and scripts regenerate all figures and tables from those files. The supplementary material holds the code, every data builder (the synthetic sets regenerate byte-for-byte from a seed) and all result files; trained checkpoints are not included in the archive. The real corpora are public (PAMAP2 and OPPORTUNITY from the UCI repository, Aria Everyday Activities under its research licence) and the preprocessing scripts are included. Training: AdamW, learning rate 10^{-4}, weight decay 0.01, batch size 2, $d_{\text{model}} = 256$, 256 slots, 64 tokens, store-supervision weight 1.0 annealed to zero; epochs and seeds are stated per table. One run uses one 80 GB GPU (NVIDIA H100): about 40 minutes for PAMAP2, 2 hours for OPPORTUNITY, 1.5 hours for the diagnostic set and 3 hours for Aria. Seeds, splits (by recording or by subject, never by question), placeholder token ids and the two scoring rules are fixed in code; the published benchmarks are scored by their own released code.

Ethics statement

All corpora are public and de-identified: PAMAP2 and OPPORTUNITY are released for research by their authors, Aria Everyday Activities under Meta's research licence with consenting wearers, PEERS and CHB-MIT with the consent procedures of their original studies. No new data was collected. A bounded memory over physiological and egocentric recordings could support continuous

monitoring of people; the models here are trained on those corpora only, answer questions about recordings the user supplies, and nothing released identifies a wearer. We see no other ethical concern specific to this work.

AI USE STATEMENT

In this work, we used generative AI tools for writing and debugging experiment code, for drafting and editing prose, and for literature search. We have not used generative AI tools to generate research ideas, to fabricate or impute experimental data, or to produce reviews, and the remaining disclosure categories are not applicable to this work. We have reviewed all AI-assisted work: the tool was an LLM coding assistant used interactively by the authors; every measurement is read from a result file written by the run that produced it and was re-derived by the authors; AI-generated code was tested against held-out checks (probes, majority baselines, cross-scoring); every citation was checked by the authors against the cited source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Emran Ali, Maia Angelova, and Chandan Karmakar. Epileptic seizure detection using chb-mit dataset: The overlooked perspectives. *Royal Society open science*, 11(5):230601, 2024.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8GjSf9Rh7Z.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, 2026.

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ayi7qezU87.

Vivek Chari, Guanghui Qin, and Benjamin Van Durme. Kv-distill: Nearly lossless learnable context compression for llms. *arXiv preprint arXiv:2503.10337*, 2025.

Chunliang Chen, Ming Guan, Xiao Lin, Jiaxu Li, Luxi Lin, Qiyi Wang, Xiangyu Chen, Jixiang Luo, Changzhi Sun, Dell Zhang, et al. Telemem: Building long-term and multimodal memory for agentic ai. *arXiv preprint arXiv:2601.06037*, 2025a.

Yijing Chen, Yihan Wu, Kaisi Guan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. Chronusomni: Improving time awareness of omni large language models. *arXiv preprint arXiv:2512.09841*, 2025b.

Yuliang Chen, Arvind Pillai, Yu Wu, Tess Z Griffin, Lisa Marsch, Michael V. Heinz, Nicholas C. Jacobson, and Andrew Campbell. Learning transferable sensor models via language-informed pretraining. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=0Yh94wslQN.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, 2023.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Zican Dong, Peiyu Liu, Junyi Li, Zhipeng Chen, Han Peng, Shuo Wang, and Xin Zhao. ForesightKV: Optimizing KV cache eviction for reasoning models by learning long-term contribution. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=znV8JHv8b8.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z Pan. Rethinking memory in llm based agents: Representations, operations, and emerging topics. *arXiv preprint arXiv:2505.00675*, 2025.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uREj4ZuGJE.

K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. Ruiz Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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, 2024. doi: 10.1109/CVPR52733.2024.01282.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280. IEEE, 2024a.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV*, pp. 202–218, Berlin, Heidelberg, 2024b. Springer-Verlag. ISBN 978-3-031-73038-2. doi: 10.1007/978-3-031-73039-9_12. URL https://doi.org/10.1007/978-3-031-73039-9_12.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multitask foundation model for bridging the gap between language and eeg signals. In *International Conference on Learning Representations*, volume 2025, pp. 55436–55457, 2025.

Michael J Kahana, Eash V Aggarwal, and Tung D Phan. The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(12):1857, 2018.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott R. Ceresnak, Aydin Zahedivash, Alexander T Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Oppers Aalami, Filipe Barata, and Paul Schmiedmayer.

OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FHDRzhKm0f.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Yang Li, Jiaxiang Liu, Yusong Wang, Yujie Wu, and Mingkun Xu. BMAM: Brain-inspired multi-agent memory framework. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 39604–39626, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1973. URL https://aclanthology.org/2026.findings-acl.1973/.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. Sensorllm: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 354–379, 2025a.

Zongqian Li, Yixuan Su, and Nigel Collier. 500xcompressor: Generalized prompt compression for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25081–25091, 2025b.

Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang, and Shu-Tao Xia. From verbatim to gist: Distilling pyramidal multimodal memory via semantic information bottleneck for long-horizon video agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11601–11617, 2026.

Bo Liu and Qiang Liu. Maglev: Sliding recurrent memory. *arXiv preprint arXiv:2608.02870*, 2026.

Chengzhi Liu, Yuzhe YANG, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. Worldmemarena: Evaluating multimodal agent memory through action–world interaction. In *COLM 2026 The 2nd Workshop on Lifelong Agents: Learning, Aligning, and Evolving*, 2026. URL https://openreview.net/forum?id=IZK0jfy4Ev.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, et al. Memverse: Multimodal memory for lifelong learning agents. *arXiv preprint arXiv:2512.03627*, 2025.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=2DtxPCL3T5.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel eeg with time-frequency motif learning. In *International Conference on Learning Representations*, volume 2026, pp. 36611–36637, 2026.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pp. 108–109. IEEE, 2012.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, et al. Memlens: Benchmarking multimodal long-term memory in large vision-language models. *arXiv preprint arXiv:2605.14906*, 2026.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pp. 233–240. IEEE, 2010.

Shuvendu Roy, Mengyao Zhai, Hossein Hajimirsadeghi, and Golnoosh Samei. Coverage-driven kv cache eviction for efficient and improved inference of llm. *arXiv preprint arXiv:2606.29563*, 2026.

Michael S Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19070–19081. IEEE, 2023.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Nirhoshan Sivaroopan, Albert Zomaya, and Kanchana Thilakarathna. Stella: Efficient sensor-to-llm translation for on-device human activity recognition. *arXiv preprint arXiv:2607.03089*, 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.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TrjbxzRcnf-.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vGeNaFHdET.

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.

Sikuan Yan, Sicheng Dong, Haotong Wang, Ercong Nie, Yilun Liu, Jinhe Bi, Yingjie Xu, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, et al. Pyravid: Hierarchical multimodal memory for long-horizon video reasoning. *arXiv preprint arXiv:2605.17065*, 2026.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Yiqun Yao, Naitong Yu, Xiang Li, Xin Jiang, Xuezhi Fang, Wenjia Ma, Xuying Meng, Jing Li, Aixin Sun, and Yequan Wang. Egomem: Lifelong memory agent for full-duplex omnimodal models. *arXiv preprint arXiv:2509.11914*, 2025.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

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

Y. Zhang, K. Ayush, S. Qiao, A. A. Heydari, G. Narayanswamy, M. A. Xu, A. A. Metwally, S. Xu, J. Garrison, X. Xu, T. Althoff, Y. Liu, P. Kohli, J. Zhan, M. Malhotra, S. Patel, C. Mascolo, X. Liu, D. McDuff, and Y. Yang. SensorLM: Learning the language of wearable sensors. arXiv:2506.09108, 2025b.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. Belt: bootstrapped eeg-to-language training by natural language supervision. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3278–3288, 2024.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *Advances in Neural Information Processing Systems*, 38: 87150–87195, 2026.

Junyi Zou and Avrova Donz. Memory-managed long-context attention: A preliminary study of editable request-local memory. *arXiv preprint arXiv:2606.28876*, 2026.

A REAL-CORPUS TABLES

Evaluation protocol. Closed families are likelihood-ranked over each family's answer set and every number is the last training epoch, with no selection on held-out data. Every $\pm$ in the paper is a sample standard deviation: across seeds on the synthetic sets and on Aria, across held-out splits on PAMAP2 and OPPORTUNITY. *Dense-event set:* 12,000 of the 21,000 training questions, three epochs, 400 questions of the held-out subject, memory-only prompt. *Second backbone:* Qwen3.8-27B, frozen, seed 4 and the same 400 held-out questions as the first Qwen3-Omni seed; batch size 1 (the model does not fit two examples on one 80 GB GPU) and learning rate 5×10^{-5}, half the rate used at batch size 2. At 10^{-4} the unbound run ends at 38.8% (`when` 7%, `first` 96%) and the

bound run, which tracks it for 10,500 steps, then diverges and ends at 16.2%; we report that run as a training failure and not as evidence about the method. *PAMAP2:* each seed holds out a different pair of subjects and every row scores every question of that pair (894–1,034 per split), so each bound memory is paired with its unbound counterpart; the per-split gains are $+11.1/+5.3/+2.5$ (label-free store), $+10.7/+7.0/+8.6$ (labelled store), $+4.2/+10.2/+8.8$ (question-blind store) and $+0.8/+16.8/+6.8$ (mean-pool). In the symbol-count sweep of Fig. 3 the bound points score the first 400 questions of the first split and the unbound point every question. *OPPORTUNITY:* rows above the rule except the label-free store, and the mean-pool with symbols, score every question of the held-out subject (927–1,263); the label-free store, the question-blind store and the flat tokenizer with symbols score its first 400; each pair across the rule (mean-pool, label-free store, question-blind store) uses identical questions. *Aria:* five locations, one held out; every seed holds out the same one, so ± is across training seeds. The full build has 6,151 questions (1,378 held out; majority 11.3% on `when`); the three-seed diagnosis row was trained on an earlier build (1,284 questions, 253 held out; majority 14.1% on `when`, 42.6% on `place`). All runs are four epochs.

	state (13-way)	first	after	count	before/after	all
	balanced accuracy (%)					acc. (%)
chance (1/classes present)	9.6	11.1	12.5	33.3	50.0	–
perception→text: classifier transcript, 2 seeds	35.4±6.5	62.4±1.7	39.6±3.2	51.7±3.0	62.1±4.3	40.1±3.2
+ interval of the queried step given, 3 seeds	44.1±6.1	60.4±3.5	37.9±3.7	47.7±7.2	61.7±3.1	43.8±4.0
perception→text: ground-truth transcript, 2 seeds	38.2±4.5	89.1±0.5	59.5±2.4	78.8±1.9	67.8±1.1	53.1±0.3
+ interval of the queried step given, 3 seeds	52.7±2.7	88.3±1.3	58.6±2.3	77.2±3.1	66.8±2.0	60.5±0.6
flat sensor tokenizer, 3 seeds	54.2±5.7	55.7±5.1	92.5±1.1	49.7±0.6	89.7±3.8	61.5±0.6
bucketed mean-pool, no learned store, 3 seeds	61.5±6.5	66.2±2.2	90.8±6.7	60.6±7.2	90.8±5.0	68.1±5.4
per-stream stores, time-aligned read, 3 seeds	57.6±4.1	63.2±4.5	86.8±6.3	63.1±11.5	90.6±1.8	66.6±3.6
shared store, no store supervision, 3 seeds	56.4±5.2	64.1±7.9	87.0±11.8	56.1±3.4	82.3±10.9	65.1±4.8
shared store, label-free masked supervision, 3 seeds	62.5±6.5	65.5±6.3	92.7±5.8	56.6±0.6	88.9±1.1	68.7±3.0
UMMI (shared store + supervision), 3 seeds	62.5±7.4	67.0±6.4	88.5±9.1	60.3±1.3	86.3±1.7	68.3±3.7
flat sensor tokenizer + time symbols, 3 seeds	50.8±1.0	54.9±6.6	89.5±2.1	53.5±4.8	82.8±5.0	60.5±1.9
bucketed mean-pool + time symbols, 3 seeds	72.4±4.7	69.8±9.2	89.8±3.1	**63.9±6.5**	92.0±3.0	76.2±4.5
question-blind store (no sharing), 3 seeds	61.3±3.8	68.2±3.2	88.2±2.8	58.9±2.0	93.4±3.1	67.9±0.8
+ time symbols, 3 seeds	73.2±4.8	67.8±10.2	88.3±8.5	58.5±2.1	94.0±1.7	75.6±3.3
UMMI, label-free + time symbols, 3 seeds	71.1±5.9	69.0±7.1	**93.3±2.8**	56.7±1.7	92.8±2.8	74.9±3.3
UMMI (shared store + supervision) + time symbols, 3 seeds	**74.5±1.9**	**70.2±6.4**	88.6±5.2	61.2±5.3	**94.5±1.6**	**77.1±2.8**

Table 3: PAMAP2, all configurations (the complete version of Table 2; same protocol). Per-split gains of the bound memories are listed in the evaluation protocol above.

	state (5-way)	first	after	count	before/after	all
	balanced accuracy (%)					acc. (%)
chance (1/classes present)	20.0	25.0	25.0	25.0	50.0	–
perception→text: classifier transcript, 2 seeds	27.3±4.0	67.2±6.7	20.0±0.9	33.6±2.4	51.5±0.1	41.1±4.7
+ interval of the queried step given, 3 seeds	42.7±3.5	65.8±5.3	19.8±0.7	34.4±2.1	51.1±0.7	43.3±3.2
perception→text: ground-truth transcript, 2 seeds	32.9±8.1	80.1±0.7	41.2±17.9	63.2±4.7	52.0±1.2	46.5±2.3
+ interval of the queried step given, 3 seeds	57.3±2.3	81.6±2.6	42.2±12.8	63.3±3.3	52.1±0.9	53.1±1.2
flat sensor tokenizer	19.9	25.0	66.7	33.3	58.3	44.3
bucketed mean-pool, no learned store, 2 seeds	23.8±3.0	50.5±11.0	45.5±15.8	40.0±0.1	71.1±8.2	52.0±3.5
UMMI (shared store + supervision), 2 seeds	25.0±1.5	46.3±2.2	50.3±4.4	49.4±18.7	74.1±1.6	52.9±1.5
shared store, label-free supervision, 3 seeds	19.2±1.1	36.7±13.2	39.6±26.3	39.0±17.6	58.3±7.5	41.6±3.8
flat sensor tokenizer + time symbols	17.4	25.0	51.3	33.3	48.6	37.0
shared store, label-free + time symbols, 3 seeds	33.3±7.5	44.1±7.4	52.2±10.2	43.2±17.7	65.2±8.7	53.8±7.9
bucketed mean-pool + time symbols, 3 seeds (all questions)	43.0±5.3	52.2±10.3	53.6±3.1	43.3±11.7	**78.0±1.1**	**63.6±1.9**
question-blind store, no symbols, seed 0	20.9	49.3	59.3	42.6	69.2	46.5
question-blind store + time symbols, 3 seeds	**43.3±8.5**	**61.1±6.6**	**55.6±6.8**	**49.1±13.2**	77.3±1.0	62.5±2.7

Table 4: OPPORTUNITY, as Table 2: one held-out subject per seed, last epoch of four; rows below the rule add the sixteen time symbols. Each pair across the rule is scored on identical questions (Appendix A).

B PUBLISHED BENCHMARKS: THE COST OF DISTRACTORS, ISOLATED

Three published text benchmarks let us test, with the frozen backbone alone, the claim behind a bounded memory: that a budget spent well beats an unbounded context. MemLens (Ren et al., 2026) (NeurIPS 2026 D&B) and LongMemEval (Wu et al., 2025) (ICLR 2025) ship an identical record schema; LoCoMo (Maharana et al., 2024) (ACL 2024) is converted to it. All three are scored by MemLens's own code and, where reported, judged by the same served model; three judges on identical predictions spanned 9.5%, so every judged number carries its judge and comparisons stay within one.

	when	first	after	turns	before/after	place	all
majority, full build	11.3	0.9	0.5	26.7	50.8	50.0	–
timestamped transcript in the prompt, zero-shot (no memory)	**77.0**	**89.4**	**78.1**	**40.1**	51.7	32.2	**59.9**
shared store + text stream, label-free, full build	14.2	61.1	54.3	18.8	48.3	50.0	37.5
+ time symbols (8)	17.6	65.5	51.9	36.1	50.8	50.0	41.4
flat tokenizer, full build	13.9	62.0	46.2	11.4	48.3	50.0	35.2
+ time symbols (8)	15.0	65.5	52.4	11.4	44.5	50.0	36.1
shared store + text stream, label-free, earlier build, 3 seeds	19.3±2.4	87.7±3.0	69.3±6.1	20.7±10.2	50.4±1.4	42.6±0.0	42.2±0.8

Table 5: Aria Everyday Activities, accuracy per family on the held-out location, memory-only prompt. The transcript row and the four middle rows score the same 1,378 held-out questions, so each "+ time symbols" row is paired with the row above it; the last row is the three-seed diagnosis of §4 on an earlier build of the corpus (Appendix A).

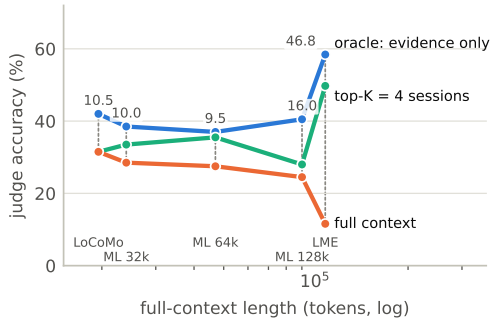

Figure 4: Five settings ordered by full-context length. Full and oracle hold identical evidence; their gap (dotted, labelled, in points of judge accuracy) is the cost of distractors: about ten points across the three in-window settings and larger beyond, where the two longest full contexts exceed the backbone's window. The oracle stays within a narrow band.

Up to five conditions share each question: *blind* (no history), *recency* (the last sessions; MemLens 32k and LongMemEval only), *full* (every session), *top-K* (four sessions by BM25), and *oracle* (exactly the answer-bearing sessions, from the benchmark's labels). Fig. 4 and Table 6 give the decomposition; Table 7 lists every setting.

Three things follow. First, the cost of distractors at identical evidence is real and grows with length *inside* the window: on LongMemEval the oracle sessions padded with random distractors to 8k, 24k and 40k tokens read 57, 48 and 39 under the judge against 58 for the oracle alone (Table 6), a 20-point cost at 40k with nothing but distractors added; MemLens 128k pads from 40 to 29. Second, the catastrophic numbers are window overflow: the 107k-token LongMemEval context is $1.6\times$ the backbone's window and reads 12, close to blind; we report it as overflow, not as a distractor cost. Cells carry a standard error of about 3.5 points, so the trend, not any single step, is the finding. Third, a bounded budget spent on the right sessions is close to the ceiling: BM25's four sessions take 85% of what the oracle reaches on LongMemEval, recency alone is no better than nothing (10.5 vs 9.7), and a learned selector has 5–10 points to win, which is why we did not train one.

C THE CHAIN BENCHMARK IN FULL

Signal parameters. An engaged cue adds a 0.8-amplitude sinusoidal burst (± 6 steps) to a band-limited EEG oscillation (amplitude 0.15, noise 0.12); a firing muscle adds a 0.9-amplitude burst to Gaussian EMG noise (0.08); LiDAR is a piecewise-constant range stepping to 0.15 (near) or 0.9 (far) at the cue.

The facts are recoverable; the model does not receive them. With the flat tokenizer, depth-4 sits at 8% and each sensor fact at its majority class (engaged 52% vs. 49%, muscle 55% vs. 48%, near 55%); only colour, carried by the visual tower, moves ($26 \rightarrow 44\%$). The probe of §3, given the cue position, reads the same facts at 100% from the same window. What is lost is lost between the window and the language model.

	oracle	padded 8k	padded 24k	padded 40k	full
LongMemEval	4k: 48.4 / 58.4	8k: 48.7 / 57.4	23k: 45.1 / 47.7	39k: 40.5 / 38.5	107k†: 10.0 / 11.6
MemLens 128k	8k: 38.5 / 40.5	9k: 37.5 / 34.5	23k: 29.5 / 33.0	39k: 32.0 / 29.0	90k†: 22.5 / 24.5

Table 6: Length varied inside the window at identical evidence: oracle sessions plus random distractor sessions up to N tokens (SubEM / judge, %); †beyond the 65k window. Blind scores 8–13 under the judge.

		judge accuracy (%)			distractor cost (oracle − full)		
setting	ctx tokens	full	top-K	oracle	judge	SubEM	F1
LoCoMo	19,492	31.5	31.5	42.0	**10.5**	5.0	6.3
MemLens 32k	24,032	28.5	33.5	38.5	**10.0**	8.5	8.7
MemLens 64k	46,973	27.5	35.5	37.0	**9.5**	9.5	12.8
MemLens 128k	90,098	24.5	28.0	40.5	**16.0**	16.0	11.8
LongMemEval	107,302	11.6	49.7	58.4	**46.8**	38.4	28.8

Table 7: Five settings, 190–200 stratified questions each, one served judge (Qwen3.8-27B with MemLens's judge prompt) for every cell. Distractor cost is oracle − full at identical evidence, in points, under the judge and under MemLens's SubEM and F1. The two longest full contexts (90k and 107k tokens) exceed the backbone's 65,536-token window, so their costs include window overflow; Table 6 varies length inside the window.

Training the backbone does not close the gap. Rank-32 LoRA on the thinker for three epochs leaves depth-4 at 5%. On an earlier 96-step version of this benchmark, four LoRA conditions (rank 16, rank 64, joint, and LoRA on a frozen pre-trained interface) span 0.5% in best accuracy while one condition swings 6.5% between its own epochs and seed spread is 4.4%; more trainable backbone parameters correlate with less stability, not more accuracy. Told all four facts in words, the frozen backbone applies the rule at 64% under likelihood-ranked scoring (Appendix F), so the reasoning is present. The bottleneck is retrieval, and it compounds: the product of the four single-fact accuracies, $0.44 \times 0.52 \times 0.55 \times 0.55 = 6.9\%$, predicts the measured depth-4 of 8%. We do not call this a law—the literature reports and our data agree that measured multi-hop accuracy exceeds the product—but it identifies the only lever: the single-fact terms.

Conditioning the read on the question does not help. A question vector added to the read queries moves the closed-set families by $+0.0$ to $+3.0$ points against a matched control, inside seed spread, and leaves the sensor facts at their floors. Reading cannot recover what the store never kept.

Address the store by time. A store-supervision head that predicts, for each of sixteen time buckets, whether a cue occurred and its four fields was trained on the store with the language model removed entirely. With free learned queries reading the store, every field sat exactly on its floor for nine epochs (presence 62.5%, colour 25%, binaries 50%). With a read in which bucket i pools the slots covering interval i, the same store reaches engaged 99.8% and muscle 100%—the probe's ceiling. Free learned queries lose to structural time alignment here and again in the question-conditioned read above; the design is known (Ryoo et al., 2023; Liu & Liu, 2026); the cost of its absence was not measured. Trained jointly, the store supervision is dose-dependent (depth-4 $6/33/48\%$ at weight $0/0.1/1$; Fig. 6) and annealing it to zero over training matches the constant-weight result, so it is scaffolding rather than part of the objective; five seeds give $44.4 \pm 8.8\%$ (Fig. 5). Colour, which does not pass through the sensor store, is unchanged in both arms—the internal control that the effect is not leakage.

The visual tower destroys time. With the shared store and the whole image, colour plateaued at 44% while the sensor facts read through the store reached 90%. Recolouring one strip cell at a time and recording which pooled visual token moves shows why: sixteen time-ordered cells touch only seven tokens, and cells 6–13 collapse onto one. The audio tower probed the same way is exact (sixteen bursts, sixteen embeddings, mean error 0.3). The visual tower is a spatial encoder;

	colour (%)	depth-2 (%)	depth-4 (%)	open F1 (%)
UMMI (full), batch 2, 5 seeds	98.6±1.9	97.8±3.4	96.0±4.2	37.8±25.9
− sharing (per-stream stores, time-aligned read)	44.0	86.0	42.0	31.9
− time-aligned read (attention read), 3 seeds	42.7±5.7	44.0±4.3	16.0±2.8	39.5±29.8
− store supervision, 3 seeds	41.3±2.5	28.7±4.1	7.3±1.9	22.6±22.0
− time-slicing (shared store, whole image), 3 seeds	44.0±0.8	92.7±6.8	54.0±2.9	32.4±15.2
bucketed mean-pool, no learned store, 3 seeds	83.3±13.6	79.3±23.6	76.0±22.5	11.4±2.1
flat tokenizer, 512 tokens (uncompressed)	44.0	18.0	6.0	0.4
UMMI, memory-only prompt (no tower features in context), 2 seeds	96.0±0.0	98.0±0.0	93.0±7.1	6.3±3.3
UMMI, label-free masked store supervision	96.0	88.0	90.0	39.1
UMMI, memory-only prompt + time symbols	100.0	100.0	100.0	18.3

Table 8: Leave-one-out ablation on the chain task, last epoch of three, seed 4 unless a seed count is given (mean ± s.d.). Single-seed rows are point estimates; the five-seed row puts the closed families within ± 4. The open family (every cue's time, colour and outcome, scored by time-tolerant F1 against a blind 31%) remains unstable across seeds and is the metric we do not claim.

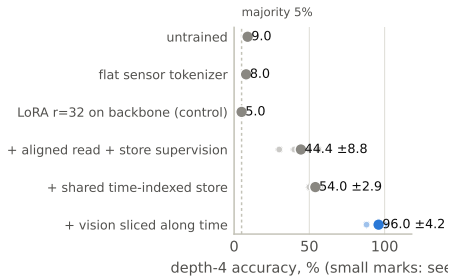

Figure 5: Depth-4 accuracy (16-way, majority 5%) on the chain task as components are added; each mark is a seed, the large mark the mean. A flat tokenizer and a tuned backbone (a control) sit at the floor; five seeds of the full store give $96.0 \pm 4.2\%$ at the last epoch.

a temporally structured image is mishandled by construction. Cutting the strip into sixteen slices along time before the tower restores a one-to-one correspondence (verified: every cell moves its own block, in order); scattering the slices onto the store's time grid—prompt byte-identical, only the store's contents change—takes colour from $44.0\pm0.8\%$ to $98.6\pm1.9\%$ and depth-4 from $54.0\pm2.9\%$ (three seeds) to $96.0 \pm 4.2\%$ (five seeds; Table 8). A shared time-indexed store into which every stream writes at its own rate is what these two measurements amount to, and it is what neither the single-stream memories nor the sensor encoders of §2 do.

Ablation. The ladder of Fig. 5 adds parts cumulatively; Table 8 removes each from the full model, one at a time, at the same seed and budget, and adds two baselines the store must beat: a bucketed mean-pool of tokenizer output with no learned slots, and the flat tokenizer given 512 tokens, one per step, so that nothing is compressed. Three parts are load-bearing: reading by attention instead of by time costs 80 points of depth-4 ($96 \to 16$, three seeds), removing store supervision costs 89 ($96 \to 7$, three seeds), and giving each stream its own time-indexed store, with the same time-aligned read and sliced vision, costs 54 ($96 \to 42$) with colour back at 44%—sliced visual tokens help only when they land in the *same* store as the sensor tokens. Annealing, halving the slices and quartering the slots cost nothing on the closed families (the no-annealing run loses the open family, 6% F1, inside that family's variance), and accuracy is flat across cue positions ($100/99/100/93/97/98\%$ from first to sixth): an early event costs no more to recall than a late one. The mean-pool baseline is the honest comparison: bucketing tokenizer output by time with no learned store reaches $76 \pm 22\%$ depth-4 over three seeds (50, 88, 90) against the store's 96 ± 4; indexing by time does most of the work, and the learned slots make it reliable. Finally, the store is the only path the backbone needs: trained and evaluated with *no* tower features in the prompt—the 64 memory tokens are all the model sees of the recording—depth-4 is 88 and 98% on two seeds against 82% for the mean-pool under the same constraint, and supervising the store with no labels at all, by predicting masked buckets from the others, keeps depth-4 at 90% and colour at 96%. The numbers above do not lean on raw audio or vision being present at question time, nor on field labels. The uncompressed baseline settles the

family	asks for	classes	majority
`fact`$_{c/e/m/n}$	one field of cue k (colour / engaged / muscle / near)	4 / 2 / 2 / 2	26 / 49 / 48 / 55
`depth-2`	engaged $\wedge$ muscle at cue k	4	23
`depth-4`	the rule's outcome at cue k	16	5
`open`	every cue: time, colour, outcome	open	—

Table 9: Question families over one recording. The open family has no answer set (3,000 references, all distinct) and is scored by time-tolerant matching per field against a blind baseline that knows only the format and the marginals (F1 31.3%, colour 28%, outcome 7.2%).

other direction: the flat tokenizer given 512 tokens, one per step, sits at chance on depth-4 (6%) and at 44% colour, no better than with 64. Budget is not the constraint; addressing is.

The backbone reasons over the memory: rules it has never seen. Two memories trained under one fixed rule, evaluated on recordings whose prompt states a *different* rule table (all sixteen outcomes remapped, signals unchanged), keep every fact family at 94–100% and depth-2 at 94–98% but drop depth-4 to 14–18% (chance 6%): under a fixed rule the mapping is absorbed in training. Trained instead with a different rule per recording, so that the only way to answer is to read the rule in the prompt and apply it to the four recalled fields, the same memory on the same frozen backbone reaches 96% and 100% depth-4 on a rule it has never seen (two seeds, six epochs; colour and depth-2 at 100%), and 100% on every family when the memory is the model's *only* access to the recording (memory-only prompt, seed 4). One of the seeds loses the open family (1–3% F1), the instability of §7 again. A control fixes what this shows: the same four fields written in *words* plus a 64-embedding soft prefix (Lester et al., 2021) trained under the same mixed-rule regime also applies never-seen rules at 100%, while the untouched backbone given the words reads 8% under this prompt (64% under another, Appendix F). Rule-following is installed by sixty-four trained embeddings, prefix or memory. The rule swap thus shows that the store does not memorise the mapping and delivers the fields at 100%; it does not show reasoning the backbone could not otherwise be taught.

The delivered model speaks. The frozen talker and code2wav sit downstream of the thinker, whose input carries the memory tokens, so the memory reaches speech with no retraining. Loading the full model with the seed-4 memory and asking four held-out depth-4 questions, the model says "reflex", "flinch", "observe", "grasp"—each the reference—one token apiece, and again on PAMAP2 (§8).

What is not solved. The open-ended family is the metric a memory model should be judged on and it remains unstable: five seeds give outcome $61 \pm 27.6\%$ (three at 83%, two collapsed near 25%) against a blind 7.2%; timing F1 rose from below-blind (25.9) to 56.8% once the store carried a shared time index. The closed families explain the open result; they do not stand in for it.

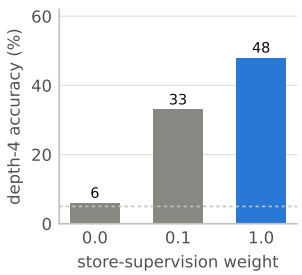

Figure 6: Depth-4 accuracy against the store-supervision weight (per-stream store, whole image, seed 4). Monotone; the control at weight 0 is the same network trained on answer tokens alone.

D TEMPORAL FIDELITY

Inject an isolated event at fraction t of an encoder's input, find the output position whose embedding changes most, and score $1 - \text{mean}\,|\text{observed} - \text{expected}|/L$ over sixteen probes, L being the encoder's number of output positions (320 audio embeddings, 16 cells for the pooled visual tower, 128 tokens for the sliced one, 512 steps for the sensor path). Exact is 1.0; an encoder that routes every moment to the same output sits near 0.75 or below. Table 10 gives every encoder: audio tower 0.999, pooled visual tower 0.754 (seven distinct positions of sixteen), time-sliced visual tower 0.996, sensor path 1.000.

The measurements suggest that temporal information lost along a modality path can limit downstream recall. We treat temporal fidelity as an empirical diagnostic, rather than a formal upper bound on task accuracy. Colour, through the 0.754-fidelity path, plateaued at 44%; the sensor facts, through paths at fidelity 0.98–1.00, reached 90%; raising the visual path to 0.996 raised colour to 99% and left the sensor facts where they were. Sweeping the number of slices gives the slice-count sweep: at 1/2/4/8/16 slices depth-4 reads 62/46/86/94/96% and colour 54/54/90/98/99% (seed 4, whose one-slice colour of 54 sits above the 44 of seeds 0–2): below the tower's seven distinct positions the curve is flat and noisy, and it saturates once every cell has its own position. The asymmetry between the two backbone towers has a cause and it is not pooling in general—one tower is temporal and one is spatial—so the recipe is "encode per time interval wherever the tower does not."

encoder	fidelity	mean error	distinct / 16	monotone
backbone audio tower	0.999	0.3	16	yes
backbone visual tower, pooled	**0.754**	**3.9**	**7**	**no**
backbone visual tower, 16 time slices	0.996	0.5	16	yes
sensor field / store / prompt tokens	1.000 / 1.000 / 0.984	$\leq$1.0	16	yes

Table 10: Temporal fidelity of every encoder. Exactly one path is low, and it is the path that carried the one stuck fact.

E NEGATIVE RESULTS IN DETAIL

An event-indexed store. Before the diagnosis of §4 we tried to fix the dense-item failure in the store: slots following *events* rather than uniform intervals, with change points proposed by a small causal predictor's next-step error across streams (surprise, z-scored causally with a warm-up mask), K events per recording, non-maximum suppression at half an interval, read in event order or binned back onto the time grid. It did not help anywhere: PAMAP2 66.8% against 66.9 for the interval store on the same split; Aria 42.0 (K=64), 39.0 (K=32) and 40.0 (uniform events) against 43.0 for the interval store at a one-epoch screening budget; on the dense-event set 21% against the interval store's 19 at that budget. Segment means lose item identity exactly as interval means do, and where the time reaches the model is unchanged.

A learned time marker instead of a symbol. The control that separates "the memory lacked any time marker" from "it lacked a marker the frozen model already reads" is sixteen trained vectors, one per interval, added to the memory tokens of that interval (the positional-embedding-on-memory-rows design of Ryoo et al. (2023)), with the prompt unchanged and everything else as in the symbol runs. It helps much less than the symbols. On the dense-event set the question-blind store reads 69.1$\pm$11.4% with the learned markers (56.0/77.2/74.0 over three seeds) against 67.6$\pm$12.0 without any marker and 89.1 $\pm$ 0.6 with the symbols; the interval question reads 71 $\pm$ 23 against 61 $\pm$ 24 and 98 $\pm$ 2. On PAMAP2, on identical questions, the label-free shared store reads 65.8 $\pm$ 3.2 (65.5/69.1/62.7) with the markers against 68.7 $\pm$ 3.0 without and 74.9 $\pm$ 3.3 with the symbols; on OPPORTUNITY the question-blind store reads 45.8 and 49.8 with the markers, against 46.5 without any marker on the first subject and 63.2 and 64.8 with the symbols. Under this training setup, textual labels are the more effective marker. Table 11 collects the three.

A modality never trained on is not read. With LiDAR withheld from training and attached at evaluation, the other streams' facts are intact (colour 100%, depth-2 97%) and the LiDAR fact sits

all-question accuracy (%)	no marker	learned marker (16 vectors)	time symbols (16 phrases)
dense-event set, question-blind store, 3 seeds	67.6±12.0	69.1±11.4	**89.1±0.6**
`when` family	60.7±24.4	71.3±22.7	**98.0±2.0**
PAMAP2, label-free shared store, 3 splits	68.7±3.0	65.8±3.2	**74.9±3.3**
OPPORTUNITY, question-blind store, 2 subjects	46.5 (one)	47.8±2.8	**64.0±1.1**

Table 11: The learned-marker control. Same store, same continuous-memory budget (64 embeddings), same supervision and epochs; the only difference between columns is what marks an interval: nothing, sixteen trained vectors added to its tokens, or sixteen phrases in the model's vocabulary.

at chance (54 vs 50%), capping depth-4 at 58%: the store is shared, the encoder is not zero-shot, and a new sensor costs a tokenizer pass, not a backbone pass.

Real EEG. PEERS (Kahana et al., 2018) (28 subjects) and CHB-MIT (Shoeb, 2009) (24 patients) carry per-timestep state labels, so the store supervision applies. On rebalanced questions with balanced accuracy, the counting family on PEERS reaches 65.9% against 33.3% chance, with the store supervision adding 9.8 points over its control; the per-moment state family sits at chance on both corpora. A probe splits the two failures. PEERS's state is not readable from the window (37.3% balanced vs. 37.4% shuffled): the corpus, as 1 Hz band power, carries no time-localised state, and no memory can recover it. CHB-MIT's state *is* readable (69.3% vs. 50.0% shuffled, matching our own seizure AUROC of 0.845), so its chance-level memory result is a failure of the method on that corpus, not of the corpus: the auxiliary head sees a 94%-interictal label and the store learns to predict the majority.

F Further related work and negative results

Learned-query readers and bounded memories over a frozen LM (full). The free-query read that loses in §7 is the latent-bottleneck reader of Perceiver (Jaegle et al., 2021), the Q-Former of BLIP-2 (Li et al., 2023) and Flamingo's resampler: a fixed set of learned queries cross-attends into a long input. For a time-indexed store the same construction sits on the floor unless the queries are pinned to time. Memorizing Transformers (Wu et al., 2022), the Recurrent Memory Transformer (Bulatov et al., 2022), Infini-attention (Munkhdalai et al., 2024) and Titans (Behrouz et al., 2025) keep a bounded memory beside a transformer, single-stream by construction. The streaming video models MovieChat (Song et al., 2024), MA-LMM (He et al., 2024) and Flash-VStream (Zhang et al., 2025a) are the closest relatives—a bounded, time-ordered store of frame features feeding a frozen language model—and they hold one stream at one rate. Qwen-Omni's own TMRoPE aligns audio and video tokens by absolute time inside the context window; our store aligns compressed slots outside it, for streams the backbone has no encoder for. The case here is several streams at several rates in one store, which is where addressing by time stops being free.

A second, vision–language backbone. Qwen3.8-27B-Instruct (Qwen Team, 2026) has no audio tower and no speech output, so attaching the same memory to it tests whether the recipe transfers on the vision, sensor and text streams only; it is not a second omni-modal model. Under likelihood scoring, one epoch sits at the majority floor on every family with the loss still falling; six epochs reach engaged / muscle / near at 94–96%, depth-2 at 88%, depth-4 at 54% and open F1 at 59%, against 96–100%, 98%, 96% and 38% on Qwen3-Omni after three. Colour stays at chance (24%): the sensor path transfers, the time-sliced visual path through this backbone's tower does not, and depth-4 is capped by the missing field. One seed; the recipe is learnable on a second backbone and several times slower, with the time-slicing fix backbone-specific.

Scoring is a free parameter. The same frozen backbone and the same facts in words score 7.5 / 18.3 / 64.2% under first-token argmax with two prompt layouts differing by one trailing space and under length-normalised likelihood over the answer set; we report the last wherever a backbone does not answer in one token (on Qwen3-Omni the two agree, 95.4 vs. 93.1%).

Multimodal agent memory is about video, not sensors. The 2026 literature on memory for multimodal agents is substantial—WorldMemArena (Liu et al., 2026) and MemLens (Ren et al.,

2026) benchmark it, EgoMem (Yao et al., 2025) and PyraVid (Yan et al., 2026) build hierarchical stores over egocentric and long-form video, and TeleMem (Chen et al., 2025a) and MemVerse (Liu et al., 2025) combine parametric recall with retrieval and consolidation. The substrate throughout is visual: frames, clips, egocentric streams. Raw physiological and range-sensing signals are absent, and they are the harder case, because unlike a frame a 64-second EEG window has no representation the backbone was pretrained on. Porting our memory tasks onto their video substrate is the natural follow-up and we do not claim it.

Coverage-based selection does not beat relevance. Discounting sessions that repeat query terms already held scored below BM25 top-K on needle recall in seven of eight settings, and a content-diversity variant is neutral (52.7% vs 53.7%): MemLens's needles share query terms and differ in values, so query-term coverage penalises exactly the sessions needed.

Streaming under a hard budget. Presenting the recording in eight segments with a per-slot retention gate went from 23.3% to 15.3% over three epochs with the loss moving 0.011. Forgetting as an operation, rather than a question type, is not solved here.

Compressing context into learned tokens. Our sensor tokenizer produces a fixed budget of embeddings spliced into the model's input stream, which is mechanically the construction used by gist tokens (Mu et al., 2023), AutoCompressor (Chevalier et al., 2023), ICAE (Ge et al., 2024), 500xCompressor (Li et al., 2025b) and KV-Distill (Chari et al., 2025). ICAE in particular pairs a learned encoder with LoRA updates to a frozen decoder, which is our stage-two setup with a different input. We claim no novelty in the mechanism. What differs is the source and the objective: those works compress *text the model could already read*, and are measured by how little is lost against reading it uncompressed. There is no uncompressed baseline here—the model cannot read a raw EEG trace at all—so the budget is not an efficiency knob but the interface itself.

Retention under a budget is ranked by importance, and that is the part that fails. Deciding what a bounded memory keeps is the same problem as KV-cache eviction, where H2O (Zhang et al., 2023) ranks tokens by cumulative attention, SnapKV (Li et al., 2024) by pooled attention from a recent window, Scissorhands (Liu et al., 2023) by persistence of importance, and PyramidKV (Cai et al., 2025) varies the budget by layer. Agent memory systems rank the same way: BMAM (Li et al., 2026) stores an `importance` scalar per item, and consolidation-and-forgetting surveys (Du et al., 2025) treat management, not writing or reading, as the unsolved half. In our own runs a coverage-discounted selector scored below BM25 in seven of eight settings (above), because top-K over independent scores cannot express "I already hold a copy of this" or "this one changes the state relative to my last pick". Recent eviction work is moving the same way—ForesightKV (Dong et al., 2026) learns long-term contribution rather than instantaneous attention, and coverage-driven and combinatorial eviction (Roy et al., 2026) optimise the retained set jointly—which we read as independent convergence on the same diagnosis from the efficiency side.

EEG foundation models. CBraMod (Wang et al., 2025), LaBraM (Jiang et al., 2024) and CSBrain (Zhou et al., 2026) pretrain on large EEG corpora and report seizure detection on CHB-MIT. We report CHB-MIT as well but do not place our number beside theirs: their protocol uses 10-second samples of 16 raw channels at 256 Hz over 326,993 samples, against 64-second windows of six band-power channels at 1 Hz over 3,279 here. Protocol drift on this benchmark is a documented field-wide problem (Ali et al., 2024), and a comparison across that gap would measure the protocols rather than the models, so we do not attempt one. Seizure detection is in any case a perception task and orthogonal to the memory abilities studied here; we mention it only because our event encoder shares its input domain.

G PER-TYPE RESULTS ON THE PUBLISHED BENCHMARKS

Table 12 places our conditions next to numbers reported in the benchmark and memory-system papers; Table 13 gives every condition under SubEM, F1, BLEU-1 and the judge with its prompt length; Table 14 gives LongMemEval by question type and Table 15 LoCoMo by category. Two readings matter for placement (all judged numbers use a served Qwen3.8-27B judge; three judges

on identical predictions spanned 9.5 points). First, the pattern is the benchmark's own: the LongMemEval paper reports GPT-4o at 87.0 with oracle sessions and 60.6 reading the full 115k-token history, Llama-3.1-70B at 74.4 → 33.4, and Qwen2.5-7B at 28.2 → 12.8 (Wu et al., 2025); our frozen Qwen3-Omni-30B reads 58.4 → 11.6 (Table 12), the latter beyond its window. The backbone is a mid-sized open model and the paper's own long-context readers of that size land in the same place; we do not claim a competitive memory system on text, only a controlled decomposition on one. Second, where the drop lands is informative: on LongMemEval the full history loses almost everything on single-session and multi-session questions (94 → 6, 50 → 3) while knowledge-update survives best (75 → 34); on LoCoMo the adversarial category, scored by abstention, stays near its floor under every condition (12.5–17.5 by the judge for every condition that holds evidence), while single-hop is largely solved by any condition that holds the evidence (68–83).

benchmark	system	answering model	judge	acc. (%)
LongMemEval$_S$, reported by Wu et al. (2025)				
	full history (offline reading)	GPT-4o	GPT-4o	60.6
	oracle sessions	GPT-4o	GPT-4o	87.0
	full history	Llama-3.1-70B	GPT-4o	33.4
	oracle sessions	Llama-3.1-70B	GPT-4o	74.4
	ChatGPT memory	GPT-4o	GPT-4o	57.7
	Coze memory	GPT-4o	GPT-4o	33.0
LongMemEval$_S$, this paper (190–200 stratified questions)				
	full history†	Qwen3-Omni-30B	Qwen3.8-27B	11.6
	BM25 top-4 sessions	Qwen3-Omni-30B	Qwen3.8-27B	49.7
	oracle sessions	Qwen3-Omni-30B	Qwen3.8-27B	58.4
LoCoMo, reported by the Mem0 paper (Chhikara et al., 2025)				
	full context	GPT-4o-mini	LLM judge	72.9
	Mem0	GPT-4o-mini	LLM judge	66.9
	Mem0^g	GPT-4o-mini	LLM judge	68.4
	Zep	GPT-4o-mini	LLM judge	66.0
	LangMem	GPT-4o-mini	LLM judge	58.1
	OpenAI memory	GPT-4o-mini	LLM judge	52.9
LoCoMo, this paper (200 stratified questions, 40 per category)				
	full context	Qwen3-Omni-30B	Qwen3.8-27B	31.5
	BM25 top-4 sessions	Qwen3-Omni-30B	Qwen3.8-27B	31.5
	oracle sessions	Qwen3-Omni-30B	Qwen3.8-27B	42.0
MemLens 128k, reported by Ren et al. (2026)				
	full context, best LVLM (Gemini-3.1-Pro)	Gemini-3.1-Pro	Qwen3-VL-235B	52.0
MemLens 128k, this paper				
	full context†	Qwen3-Omni-30B	Qwen3.8-27B	24.5
	oracle sessions	Qwen3-Omni-30B	Qwen3.8-27B	40.5

Table 12: Published numbers next to ours. Answering models, judges, and question samples differ across blocks, so rows are comparable only within a block; we include them so the reader can place a frozen 30B open backbone, not to rank it. The pattern we measure—oracle sessions far above the full history, a small bounded budget close to the oracle—is the pattern the LongMemEval and Mem0 papers report for GPT-4o-class models.

condition	tokens	SubEM	F1	BLEU-1	judge
LoCoMo					
blind	29	10.0	7.9	5.4	8.5
full context	19,492	25.5	17.3	11.9	31.5
top-K (BM25, 4 sessions)	3,003	25.5	21.2	15.2	31.5
oracle sessions	1,138	30.5	23.6	17.4	42.0
MemLens 32k					
blind	56	13.5	7.3	3.0	12.5
recency (4 most recent sessions)	7,225	23.0	11.1	5.8	26.5
full context	24,032	29.0	11.9	7.0	28.5
top-K (BM25, 4 sessions)	8,719	31.0	14.6	8.6	33.5
oracle sessions	7,350	37.5	20.6	14.0	38.5
MemLens 64k					
full context	46,973	28.5	6.6	3.0	27.5
top-K (BM25, 4 sessions)	8,956	32.0	17.3	11.6	35.5
oracle sessions	7,558	38.0	19.4	12.8	37.0
MemLens 128k					
full context	90,098	22.5	6.8	2.9	24.5
top-K (BM25, 4 sessions)	9,103	29.5	14.7	9.1	28.0
oracle sessions	7,959	38.5	18.6	12.6	40.5
LongMemEval					
blind	36	7.7	9.2	6.2	9.7
recency (4 most recent sessions)	8,532	6.3	9.7	7.0	10.5
full context	107,302	10.0	8.2	5.0	11.6
top-K (BM25, 4 sessions)	11,354	45.1	28.4	22.2	49.7
oracle sessions	4,771	48.4	37.0	30.5	58.4

Table 13: Every condition on every setting, in %. SubEM and F1 are MemLens's scorer; BLEU-1 follows LoCoMo (unigram precision with brevity penalty on the same normalisation); judge is Qwen3.8-27B with MemLens's prompt. MemLens 64k/128k blind conditions were not run (blind does not depend on scale).

	tokens	single-user	single-asst.	preference	multi-session	temporal	knowl. update	all
blind	36	9.1	18.2	3.3	0.0	18.2	9.1	9.7
full[†]	107,302	6.2	12.5	0.0	3.1	12.5	34.4	11.6
top-K	11,354	72.7	84.9	3.3	39.4	21.2	72.7	49.7
oracle	4,771	93.8	87.5	6.7	50.0	34.4	75.0	58.4

Table 14: LongMemEval, judge accuracy (%) by question type; 30–33 questions per type.

	tokens	single-hop	multi-hop	temporal	open-domain	adversarial
blind	29	7.5 / 10.0	9.7 / 7.5	6.4 / 17.5	3.5 / 5.0	12.6 / 2.5
full	19,492	27.6 / 67.5	20.0 / 35.0	9.5 / 22.5	13.4 / 17.5	16.0 / 15.0
top-K	3,003	33.2 / 72.5	22.9 / 25.0	17.7 / 35.0	12.2 / 12.5	19.7 / 12.5
oracle	1,138	34.0 / 82.5	28.7 / 52.5	17.6 / 45.0	12.0 / 12.5	25.6 / 17.5

Table 15: LoCoMo, F1 / judge accuracy (%) by category; 40 questions per category. The LoCoMo paper reports F1; its GPT-4-turbo reader scores 32.1 overall.