SUPERMEM: OMNI-NATIVE SUPER-HYPERGRAPH MEMORY FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory lets a frozen large language model (LLM) agent answer from experience beyond its context window, increasingly omni-modal streams of text, images, audio and video. Agent memory has grown from flat stores to hypergraphs and from text to all modalities, yet hypergraph memories index only text while all-modal ones stay flat or pairwise. This leaves three challenges: modality collapse in storage, the modality gap in recall, and repetition and revision in update. We propose **SuperMem**, an omni-native super-hypergraph memory whose hyperedges bind each unit's text with its native media nodes and nest units into episodes, with no extra LLM call. Joint cross-modal recall credits each modality above its own per-query median and delivers whole units with raw media, while conservative consolidation merges repeats and supersedes revisions without deleting. On nine text, multimodal and omni-modal benchmarks, SuperMem has the best average against 13 memory baselines on each of three frozen backbones, 9.5 points above the strongest. Our software and data are publicly available[1].

1 INTRODUCTION

Long-term memory lets a large language model (LLM) agent answer from experience that outgrows its context window: it writes the incoming stream into an external store and, when a question arrives, recalls evidence for a frozen backbone (Packer et al., 2024; Chhikara et al., 2025). That experience is increasingly omni-modal: an assistant must remember photos shared over months of chat, a document agent must read long reports, and a robot must recall what it saw and heard (Wang & Chen, 2025; Long et al., 2025). As shown in Figure 1, we study *omni-modal long-term memory*: streams of units, each a text with its images, audio or video, are all written into one store before any question, so a memory must find the right episode and deliver the evidence in the modality that holds the answer.

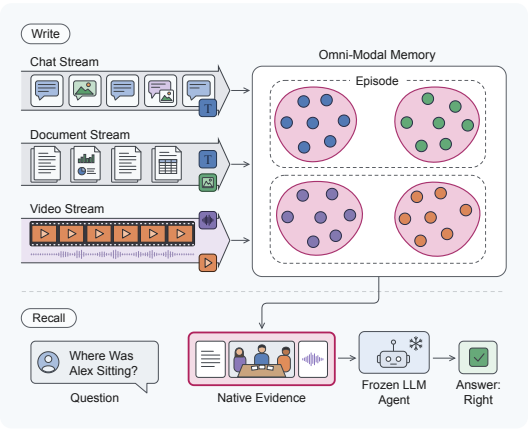


Figure 1: Omni-modal long-term memory: write every stream, then recall native evidence.

Agent memory has developed along two complementary axes. In *structure*, flat stores (Packer et al., 2024; Chhikara et al., 2025; Li et al., 2025) gave way to graphs (Rasmussen et al., 2025; Xu et al., 2025; Zhang et al., 2025) and hypergraphs (Luo et al., 2025; Zhou et al., 2026a; Yue et al., 2026a), whose one edge binds many nodes into a higher-order relation. In *modality*, text stores gave way to vision-native (Kang et al., 2026; Feng et al., 2026) and native all-modal memories (Long et al., 2025; Liu et al., 2026a) that also keep video and audio. Yet the two axes have not met: hypergraph memories index only text or keep media as captions, and all-modal memories are flat or pairwise, so the omni-modal hypergraph cell of the modality–structure grid stays empty (Figure 2).

[1] Anonymous GitHub link: `https://anonymous.4open.science/r/SuperMem`

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.

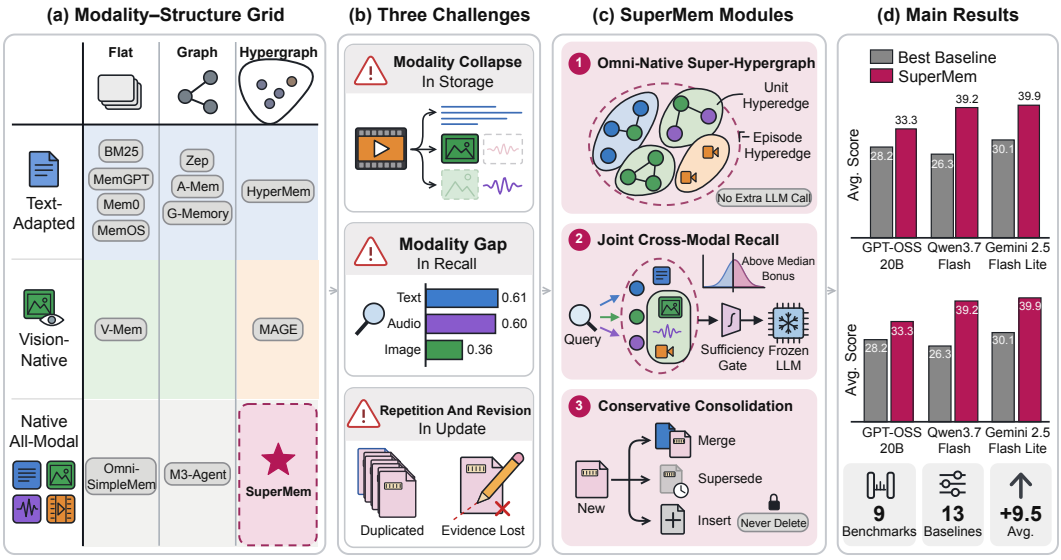


Figure 2: Overview of SuperMem: where existing memories sit on the modality–structure grid, the three challenges they leave, SuperMem's three modules, and a preview of its main results.

However, as shown in Figure 2, three challenges remain. **(i) Modality collapse in storage.** A clip binds text, frames and audio; text hypergraphs keep only its caption, and flat or pairwise ones split it. **(ii) Modality gap in recall.** A text query lies nearer other text than the image answering it (Liang et al., 2022), so fused scores rank media low. **(iii) Repetition and revision in update.** Copies crowd append-only stores, while LLM rewriting may delete evidence (Rusu et al., 2026).

To address these challenges, we propose **SuperMem** (Figure 2). Its **omni-native super-hypergraph** binds each unit's text and native media nodes in one first-level hyperedge and nests units into second-level episode hyperedges, calling no extra LLM. **Joint cross-modal recall** credits each modality only above its own per-query median and delivers whole units with raw media under a sufficiency gate. **Conservative consolidation** merges repeats and supersedes revisions by a deterministic rule that never deletes. Text-surrogate, pairwise and flat memories are provably its special cases.

We evaluate SuperMem on nine text, multimodal and omni-modal benchmarks with three frozen backbones. Against 13 memory baselines, it has the best average on every backbone, 9.5 points (34%) above the strongest, beats each in a paired sign test ($p < 10^{-8}$), on Gemini leads even at a 2k budget, and keeps recall intact under repeated writes that cost an append-only store 26.4 R@5.

2 Related Work

We summarize the two lines of agent memory that SuperMem joins; Appendix G extends this review.

Structured Agent Memory. Memory structure has grown richer, but over text alone. Beyond BM25, flat stores page a bounded context in MemGPT/Letta (Packer et al., 2024), consolidate facts by LLM in Mem0 (Chhikara et al., 2025) and rank memory by usage in MemOS (Li et al., 2025); graphs then add temporal edges in Zep/Graphiti (Rasmussen et al., 2025), triples in Mem0^g, linked notes in A-Mem (Xu et al., 2025) and insights in G-Memory (Zhang et al., 2025); and HyperMem (Yue et al., 2026a) binds n-ary facts as hyperedges. SuperMem keeps this higher-order binding but builds each hyperedge over the native embeddings of a unit's text and media, and nests the unit hyperedges of each stream into an episode hyperedge that its joint recall narrows first.

Multimodal Agent Memory. A parallel line brings memory to perception (Wang & Chen, 2025; Liu et al., 2026c) but trades native modality against structure: V-Mem (Kang et al., 2026) keeps native vectors yet routes each query to one modality, MAGE (Feng et al., 2026) builds hyperedges on textual surrogates, and Omni-SimpleMem (Liu et al., 2026a) and M3-Agent (Long et al., 2025) keep native media flat or pairwise. SuperMem keeps both, recalling each unit as one native hyperedge.

Table 1: Comparison of representative baselines, one per occupied modality–structure cell. Mem0: flat LLM-extracted facts; Zep/Graphiti: temporal text graph; HyperMem: n-ary text hyperedges; V-Mem: routed native modality lanes; MAGE: text-surrogate hyperedges; Omni-SimpleMem: flat native units; M3-Agent: pairwise text–file links; SuperMem (Ours): nested omni-native hyperedges.

Method	**Memory Organization**	**Recall & Update**
Mem0 (Chhikara et al., 2025)	$\mathcal{H} = \{c_i\}_i,\ c_i \sim \mathrm{LLM}(\tilde{c}_t)$ Flat LLM-extracted text facts.	$\mathcal{R}(q) = \text{Top-}k_{c\in\mathcal{H}} \cos(z_q, \phi(c))$ LLM decides add, update or delete.
Zep/Graphiti (Rasmussen et al., 2025)	$\mathcal{H} = (\mathcal{V}, \mathcal{E}),\ e = (v, v', c_e, \tau_e)$ Temporal graph of text facts.	$\mathcal{R}(q) = \text{Top-}k_{e\in\mathcal{E}_{\text{valid}}} \cos(z_q, \phi(c_e))$ LLM invalidates contradicted edges.
HyperMem (Yue et al., 2026a)	$\mathcal{H} = (\mathcal{V}, \mathcal{E}),\ e \in 2^{\mathcal{V}}$ n-ary text facts over entities.	$\mathcal{R}(q) = \text{Top-}k_e \max_{z\in\mathcal{Z}_e} \cos(z_q, z)$ Best of its entity and fact vectors.
V-Mem (Kang et al., 2026)	$\mathcal{H} = \{(z_t^{\mathsf{txt}}, z_t^{\mathsf{img}}, z_t^{\mathsf{aud}})\}_t$ Per-unit native modality lanes.	$\mathcal{R}(q) = \text{Top-}k_t \cos(\hat{z}_q, z_t^{\text{route}(q)})$ Query routed to one modality lane.
MAGE (Feng et al., 2026)	$e_t = \{c_t\} \cup \{\mathrm{LLM}(m)\}_{m\in\mathcal{M}_t}$ Hyperedges over textual surrogates.	$\mathcal{R}(q) = \text{Top-}1_e \cos(z_q, \phi(e))$ Media reached only through text.
Omni-SimpleMem (Liu et al., 2026a)	$\mathcal{H} = \{(\bar{c}_t, \phi(\mathcal{M}_t))\}_t$ Flat native units, novelty-gated.	$\mathcal{R}(q) = \mathrm{Dense}_k(q) \cup \mathrm{BM25}_{k'}(q)$ Hybrid candidates under a budget.
M3-Agent (Long et al., 2025)	$\mathcal{H} = \{(c_t, m) : m \in \mathcal{M}_t\}_t$ Pairwise text–file links.	$s = \frac{1}{2}[\cos(z_q, \phi(c_t)) + \cos(z_q, \phi(m))]$ Mean of text and media similarity.
SuperMem **(Ours)**	$\mathcal{H} = (\mathcal{V}, \mathcal{E}^{(1)}, \mathcal{E}^{(2)}),\ \mathcal{E}^{(2)} \subseteq 2^{\mathcal{E}^{(1)}}$ Omni-native two-level hyperedges.	$s(e) = s_q^{\mathsf{txt}}(e) + \beta_q(e)$ Joint recall; merges, never deletes.

3 METHODOLOGY: SUPERMEM

As Table 1 traces, existing memories either index text alone or keep native media flat or pairwise. SuperMem (Figure 3) addresses the three challenges of §1 with three modules, each detailed below.

Problem formulation. A benchmark $\mathcal{D}$ has streams and questions. A stream $S = (x_1, \ldots, x_{|S|})$, one episode such as a conversation, a document or a recording, is a sequence of units $x_t = (c_t, \mathcal{M}_t)$, each a text c_t with the files $\mathcal{M}_t$ it refers to, where a file m has modality $\kappa(m) \in \Sigma = \{\mathsf{txt}, \mathsf{img}, \mathsf{aud}, \mathsf{vid}\}$ (pages count as images, long videos as two-minute clips). All streams are written into one store $\mathcal{H}$ before any question $q = (c_q, \mathcal{M}_q)$ with gold answer y_q, and a frozen backbone f answers from recalled evidence $\mathcal{R}(q) = \textsc{Retrieve}(\mathcal{H}, q)$; we maximize its metric Acc:

$$\max_{\textsc{Write},\,\textsc{Retrieve}} \mathbb{E}_{(q,y_q)\sim\mathcal{D}}\big[\mathrm{Acc}(\hat{y}_q, y_q)\big], \qquad \hat{y}_q = f\big(c_q, \mathcal{M}_q, \mathcal{R}(q)\big), \qquad \mathcal{H} = \textsc{Write}(\mathcal{D}), \quad (1)$$

where WRITE adds units one by one. All rows of Table 2 share this interface and an encoder ϕ mapping every modality into $\mathbb{R}^d$ (Lee et al., 2025); cos is cosine similarity and $\mathrm{norm}(z) = z/\|z\|_2$.

3.1 OMNI-NATIVE SUPER-HYPERGRAPH MEMORY

Against modality collapse in storage, SuperMem binds native nodes in unit and episode hyperedges.

Native nodes. Each file m has one text rendering $r(m)$, a caption or transcript shared by all rows of Table 2, and $\tilde{c}_t = c_t \oplus r(\mathcal{M}_t)$ appends them to the text ($\oplus$: concatenation). A node v holds a modality $\kappa(v)$, a vector z_v and a pointer to its raw content; a unit x_t yields a unit node v_t°, passage nodes $v_{t,j}$ for the J_t passages $\pi_{t,j}$ of ℓ characters of a long $\tilde{c}_t$, and one media node v_m per file:

$$V_t = \{v_t^\circ\} \cup \{v_{t,j}\}_{j=1}^{J_t} \cup \{v_m\}_{m\in\mathcal{M}_t}, \qquad z_{v_t^\circ} = \phi(\tilde{c}_t), \quad z_{v_{t,j}} = \phi(\pi_{t,j}), \quad z_{v_m} = \phi(m). \quad (2)$$

A media node embeds the medium, not its caption, and content keys give each distinct file one node.

Two levels of hyperedges. The first-level hyperedge $e_t = V_t$ binds all nodes of one unit, of any arity. For an edge e, $\tilde{c}_e$ and $\mathcal{M}_e$ are its unit's text and files, e^μ its nodes of modality μ, $e^{\text{med}} = e \setminus e^{\mathsf{txt}}$ its media nodes and z_e° its unit-node vector. The second-level (episode) hyperedge E_S collects the edges of stream S, and both levels carry summary vectors:

$$\begin{gathered}\mathcal{H} = \big(\mathcal{V}, \mathcal{E}^{(1)}, \mathcal{E}^{(2)}\big), \quad \mathcal{V} = \textstyle\bigcup_t V_t, \quad \mathcal{E}^{(1)} = \{e_t\}_t, \quad \mathcal{E}^{(2)} = \{E_S\}_{S\in\mathcal{D}}, \quad E_S = \{e_t\}_{x_t\in S}, \\ \bar{z}_e = \mathrm{norm}\big(\mathrm{norm}(z_e^\circ) + \mathrm{norm}\big(\textstyle\sum_{v\in e^{\text{med}}} z_v\big)\big), \qquad \bar{z}_{E_S} = \mathrm{norm}\big(\textstyle\sum_{e\in E_S} \bar{z}_e\big),\end{gathered} \quad (3)$$

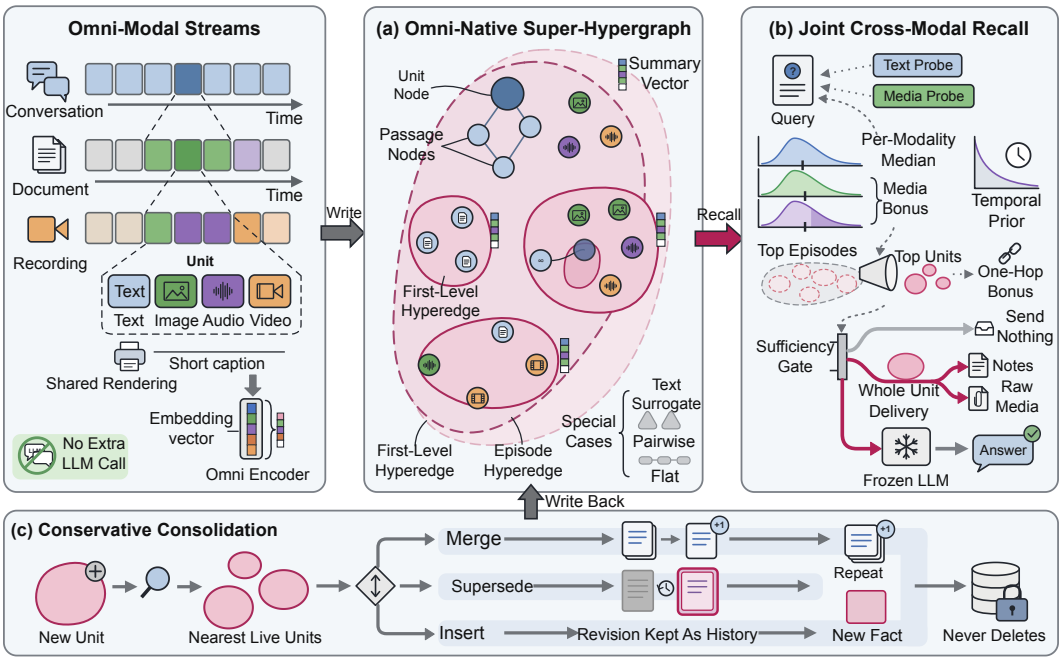


Figure 3: Methodology of SuperMem: streams are written into an omni-native super-hypergraph with no extra LLM calls, recalled jointly across modalities, and kept by conservative consolidation.

where $\bar{z}_e = \mathrm{norm}(z_e^\circ)$ without media, so the store $\mathcal{H}$ is a 2-SuperHyperGraph (Smarandache, 2020).

Cost and generality. Beyond the shared renderings, writing calls no LLM, only $1 + J_t + |\mathcal{M}_t|$ encoder calls per unit. Restricting $\mathcal{H}$ recovers the surrogate, pairwise and flat designs of Table 1:

$$e_t \mapsto e_t^{\mathsf{txt}} \ \text{(surrogate)}, \qquad e_t \mapsto \left\{e_t^{\mathsf{txt}} \cup \{v_m\}\right\}_{m \in \mathcal{M}_t} \ \text{(pairwise)}, \qquad k_2 \geq |\mathcal{E}^{(2)}| \ \text{(flat)}, \quad (4)$$

where e_t^{txt} keeps only the text nodes of e_t and k_2 is the number of episodes that recall keeps (§3.2).

Proposition 1. *Text-surrogate, pairwise and flat memories are all special cases of SuperMem.*

Proof. We provide experimental results in Sections 4 and proofs in Appendix B.1. □

3.2 Joint Cross-Modal Recall with Sufficiency-Gated Delivery

Under the modality gap (Liang et al., 2022), a text query lies nearer unrelated text than the image answering it; joint recall scores units across modalities, narrows episodes to units and gates delivery.

Joint scoring. The question probes text and media nodes alike, with $z_q = \phi(c_q)$ and, for each modality μ, the probes $P_q^\mu = \{z_q\} \cup \{\phi(m) : m \in \mathcal{M}_q, \kappa(m) = \mu\}$ that add its own files. An edge e scores a text part $s_q^{\mathsf{txt}}(e)$ plus the media bonus $\beta_q(e)$ over its media modalities Σ_e:

$$\begin{aligned} s(e) &= s_q^{\mathsf{txt}}(e) + \beta_q(e), \qquad s_q^{\mathsf{txt}}(e) = (1-\lambda_q) \max_{v \in e^{\mathsf{txt}}} \cos(z_q, z_v) + \lambda_q\, b_q(e), \\ \beta_q(e) &= \frac{w_{\mathrm{med}}}{|\Sigma_e|} \sum_{\mu \in \Sigma_e} \left[s_q^\mu(e) - \tilde{s}_q^\mu\right]_+, \qquad s_q^\mu(e) = \max_{p \in P_q^\mu,\, v \in e^\mu} \cos(p, z_v), \end{aligned} \quad (5)$$

where $b_q(e)$ is the BM25 score of c_q on $\tilde{c}_e$ normalized by its maximum (floored at b_0), $\lambda_q = w_{\mathrm{lex}}$ if any edge shares a term with c_q, else 0, w_{med} scales β_q, $\tilde{s}_q^\mu$ is the median of s_q^μ over edges holding μ, and $[x]_+ = \max(x, 0)$. The bonus thus credits each modality only above its own per-query median, so any medium counts only where it matches the question better than is typical for media of its kind.

Temporal prior. An edge has a time τ_e, its unit's date (else the last date seen in its stream) or its clip's start, and the question a time τ_q on the same clock, its date (else its stream's latest) or the

moment it asks about. Scores decay with the elapsed time, in spans of the edge's stream $S(e)$:

$$\hat{s}(e) = \rho_q(e)\, s(e), \qquad \rho_q(e) = \omega_{\text{fut}}^{\iota_q(e)}\, 2^{-[\tau_q - \tau_e]_+ / (T_{1/2}\, \text{span}(S(e)))}, \tag{6}$$

where $T_{1/2}$ is a half-life, $\iota_q(e) = \mathbb{I}[e \text{ is a clip with } \tau_e > \tau_q]$ with $\mathbb{I}$ the indicator, $\omega_{\text{fut}} < 1$ discounts such later clips, and $\rho_q(e) = 1$ if either time is unknown, so every undated unit retains its full score.

From episodes to units. An episode scores $g(E_S)$ by mixing, with weight η, its summary vector and its best member, and only members of the top k_2 episodes remain candidates $\mathcal{C}_q$. The k_1 best candidates by $\hat{s}$ are the seeds $\mathcal{K}_q^0$; a candidate sharing a node with another seed gains G-Memory's one-hop bonus δ (Zhang et al., 2025), and $\mathcal{K}_q$ keeps the k_1 best above the floor θ:

$$\begin{aligned} g(E_S) &= \eta \cos(z_q, \bar{z}_{E_S}) + (1-\eta) \max_{e \in E_S \cap \mathcal{L}_q} \hat{s}(e), \qquad \mathcal{C}_q = \mathcal{L}_q \cap \textstyle\bigcup_{E_S \in \text{Top}_{k_2}(g)} E_S, \\ s'(e) &= \hat{s}(e) + \delta\, \mathbb{I}\big[\exists\, e' \in \mathcal{K}_q^0 \setminus \{e\} :\ e \cap e' \neq \varnothing\big], \qquad \mathcal{K}_q = \text{Top}_{k_1}\big\{e \in \mathcal{C}_q :\ s'(e) \geq \theta\big\}, \end{aligned} \tag{7}$$

where $\mathcal{L}_q \subseteq \mathcal{E}^{(1)}$ are the edges eligible for q (§3.3) and Top_k keeps the k best. A kept edge returns whole, so the image beside a matched sentence reaches the backbone as well, in its own native form.

Sufficiency-gated delivery. With $s_q^\star = \max_{e \in \mathcal{K}_q} s'(e)$, or $-\infty$ if $\mathcal{K}_q = \varnothing$, the sufficiency gate sends nothing below θ_{d}, since wrong-episode evidence misleads the backbone (*Placebo*, Table 6). Otherwise notes $\mathcal{T}_q$ carry the texts $\tilde{c}_e$ of $\mathcal{K}_q$ by rank within B characters, the top n_{full} whole and the rest as their ℓ_{w}-character window densest in query terms, then files $\mathcal{F}_q$ of the edges $\mathcal{A}_q$:

$$\mathcal{R}(q) = \begin{cases} \varnothing, & s_q^\star < \theta_{\text{d}}, \\ \mathcal{T}_q \cup \mathcal{F}_q, & \text{otherwise}, \end{cases} \qquad \begin{aligned} \mathcal{F}_q &= \textsc{Fill}\big(\textstyle\bigcup_{e \in \mathcal{A}_q} \mathcal{M}_e\big), \\ \mathcal{A}_q &= \big\{e \in \mathcal{K}_q^{[n_{\text{e}}]} :\ s'(e) \geq \max(\theta_{\text{file}},\, s_q^\star - \gamma)\big\}, \end{aligned} \tag{8}$$

where $\mathcal{K}_q^{[n_{\text{e}}]}$ are the top n_{e} edges in $\mathcal{T}_q$, θ_{file} is a floor, γ a margin, and FILL adds files by rank up to a count and byte cap; an edge holding a modality outside the set Σ_f that f reads sends no raw files. Its evidence still reaches f as text, since the note already carries the shared renderings of those files.

Proposition 2. *The media bonus never lowers a score and is immune to a constant modality gap.*

Proof. We provide experimental results in Sections 4 and proofs in Appendix B.2. □

3.3 CONSERVATIVE CONSOLIDATION

Streams repeat and revise facts: copies crowd an append-only store, while LLM rewriting may delete evidence. SuperMem instead settles each write by a deterministic, LLM-free rule that never deletes.

Three cases on write. Let $\mathcal{L} \subseteq \mathcal{E}^{(1)}$ be the live edges, those not superseded. A new edge e^+ of stream S is compared with the live edges of S: its candidates $\mathcal{N}(e^+)$ are the three nearest by unit vector above θ_{dup}, and texts are compared by slots $\Psi(c)$, pairs of a measure u and a number α:

$$\mathcal{N}(e^+) = \text{Top}_3\big\{e \in E_S \cap \mathcal{L} : \cos(z_e^\circ, z_{e^+}^\circ) \geq \theta_{\text{dup}}\big\}, \quad \Psi(c) = \big\{(u, \alpha) : \text{“}\alpha\, u\text{” occurs in } c\big\}. \tag{9}$$

Texts sharing a measure *agree* ($e \sim e^+$) if every shared measure has a common value and *conflict* ($e \perp e^+$) otherwise. Exact duplicates merge; otherwise the nearest candidate meeting a case decides:

$$e^+ \mapsto \begin{cases} \textsc{merge}(e), & e^{\text{med}} = (e^+)^{\text{med}} \wedge \big[\tilde{c}_e = \tilde{c}_{e^+} \vee \big(e \in \mathcal{N}(e^+) \wedge e \sim e^+\big)\big], \\ \textsc{supersede}(e), & e \in \mathcal{N}(e^+) \wedge e \perp e^+, \\ \textsc{insert}, & \text{otherwise}, \end{cases} \tag{10}$$

where equal media nodes mean identical files. A merge only raises the write count a_e of e; an insert adds e^+ to $\mathcal{E}^{(1)}$, E_S and $\mathcal{L}$; a supersede also removes e from $\mathcal{L}$, keeping it in $\mathcal{E}^{(1)}$. No LLM is called and nothing is pruned, and a question about the past still reaches a superseded edge:

$$\mathcal{L}_q = \mathcal{L} \cup \big\{e \in \mathcal{E}^{(1)} :\ \chi(c_q) = 1\big\}, \qquad \hat{s}(e) \leftarrow \omega_{\text{sup}}\, \hat{s}(e) \quad \text{for every } e \in \mathcal{L}_q \setminus \mathcal{L}, \tag{11}$$

where $\chi(c_q) = 1$ on a temporal cue such as *before*, and $\omega_{\text{sup}} < 1$ ranks superseded below live edges.

Proposition 3. *Consolidation never deletes, and only repeated or revised facts change recall.*

Proof. We provide experimental results in Sections 4 and proofs in Appendix B.3. □

Table 2: **Main results** on nine benchmarks with three frozen backbones (§4.1). Chips mark the modality of a store (T text-adapted, M vision-native, O native all-modal) and its structure (flat, graph, hypergraph). Subscript: gap to No memory; **bold**: best, underline: second-best memory; shaded: evidence reaches f as $r(m)$; †: reimplemented; **Avg.**: mean score over the nine benchmarks.

Model	Method	A Text			B MM			C Omni			Avg.
		LoCoMo F1	LMEval Acc	MemAB Acc	MemGal F1	MemEye Acc	LDocURL F1	M3B-r Acc	OmniVB Acc	LVOmni Acc	mean
GPT-OSS 20B	No memory	6.1	5.5	9.4	12.3	32.0	16.6	3.9	28.1	25.8	15.5
$0.03/$0.13 per 1M	BM25†	34.5$_{+28.4}$	21.1$_{+15.6}$	22.7$_{+13.3}$	18.6$_{+6.3}$	36.7$_{+4.7}$	22.7$_{+6.1}$	20.3$_{+16.4}$	23.4$_{-4.7}$	29.7$_{+3.9}$	25.5$_{+10.0}$
128K ctx · Aug 2025	MemGPT / Letta	30.3$_{+24.2}$	18.8$_{+13.3}$	32.8$_{+23.4}$	34.7$_{+22.4}$	40.6$_{+8.6}$	22.0$_{+5.4}$	17.2$_{+13.3}$	21.9$_{-6.2}$	22.7$_{-3.1}$	26.8$_{+11.3}$
Text native	Mem0	26.6$_{+20.5}$	20.3$_{+14.8}$	19.5$_{+10.1}$	20.7$_{+8.4}$	36.7$_{+4.7}$	19.8$_{+3.2}$	18.8$_{+14.9}$	23.4$_{-4.7}$	24.2$_{-1.6}$	23.3$_{+7.8}$
MM caption	Zep / Graphiti	23.2$_{+17.1}$	11.7$_{+6.2}$	15.6$_{+6.2}$	18.3$_{+6.0}$	32.8$_{+0.8}$	14.4$_{-2.2}$	9.4$_{+5.5}$	21.9$_{-6.2}$	23.4$_{-2.4}$	19.0$_{+3.4}$
Doc text	A-Mem	13.5$_{+7.4}$	8.6$_{+3.1}$	18.8$_{+9.4}$	15.5$_{+3.2}$	34.4$_{+2.4}$	18.8$_{+2.2}$	9.4$_{+5.5}$	23.4$_{-4.7}$	22.7$_{-3.1}$	18.3$_{+2.8}$
Omni → T frames	Mem0^g	26.7$_{+20.6}$	15.6$_{+10.1}$	19.5$_{+10.1}$	17.6$_{+5.3}$	32.8$_{+0.8}$	21.5$_{+4.9}$	17.2$_{+13.3}$	19.5$_{-8.6}$	28.9$_{+3.1}$	22.1$_{+6.6}$
	G-Memory	24.0$_{+17.9}$	12.5$_{+7.0}$	21.1$_{+11.7}$	27.2$_{+14.9}$	41.4$_{+9.4}$	19.7$_{+3.1}$	16.4$_{+12.5}$	26.6$_{-1.5}$	25.0$_{-0.8}$	23.8$_{+8.2}$
	MemOS	27.2$_{+21.1}$	17.2$_{+11.7}$	27.3$_{+17.9}$	30.0$_{+17.7}$	42.2$_{+10.2}$	24.3$_{+7.7}$	15.6$_{+11.7}$	24.2$_{-3.9}$	28.1$_{+2.3}$	26.2$_{+10.7}$
	HyperMem†	27.6$_{+21.5}$	20.3$_{+14.8}$	23.4$_{+14.0}$	18.7$_{+6.4}$	32.8$_{+0.8}$	22.9$_{+6.3}$	17.2$_{+13.3}$	21.1$_{-7.0}$	22.7$_{-3.1}$	23.0$_{+7.4}$
	V-Mem	35.5$_{+29.4}$	21.1$_{+15.6}$	**34.4**$_{+25.0}$	35.4$_{+23.1}$	43.8$_{+11.8}$	15.1$_{-1.5}$	22.7$_{+18.8}$	22.7$_{-5.4}$	22.7$_{-3.1}$	28.2$_{+12.6}$
	MAGE	27.0$_{+20.9}$	12.5$_{+7.0}$	22.7$_{+13.3}$	34.8$_{+22.5}$	41.4$_{+9.4}$	20.8$_{+4.2}$	17.2$_{+13.3}$	25.0$_{-3.1}$	25.0$_{-0.8}$	25.2$_{+9.6}$
	Omni-SimpleMem	31.1$_{+25.0}$	14.1$_{+8.6}$	29.7$_{+20.3}$	25.4$_{+13.1}$	39.1$_{+7.1}$	17.4$_{+0.8}$	20.3$_{+16.4}$	25.8$_{-2.3}$	25.0$_{-0.8}$	25.3$_{+9.8}$
	M3-Agent	24.6$_{+18.5}$	14.1$_{+8.6}$	20.3$_{+10.9}$	22.6$_{+10.3}$	33.6$_{+1.6}$	**25.2**$_{+8.6}$	17.2$_{+13.3}$	25.0$_{-3.1}$	25.8$_{+0.0}$	23.2$_{+7.6}$
	SuperMem	**37.8**$_{+31.7}$	**23.4**$_{+17.9}$	31.2$_{+21.8}$	**48.3**$_{+36.0}$	**49.2**$_{+17.2}$	25.1$_{+8.5}$	**23.4**$_{+19.5}$	**30.5**$_{+2.4}$	**31.2**$_{+5.4}$	**33.3**$_{+17.8}$
Qwen3.7 Flash	No memory	18.0	8.6	4.7	17.1	24.2	10.8	0.8	16.4	29.7	14.5
$0.03/$0.13 per 1M	BM25†	44.8$_{+26.8}$	22.7$_{+14.1}$	15.6$_{+10.9}$	25.0$_{+7.9}$	28.9$_{+4.7}$	14.5$_{+3.7}$	18.8$_{+18.0}$	10.2$_{-6.2}$	25.8$_{-3.9}$	22.9$_{+8.4}$
1M ctx · Jul 2026	MemGPT / Letta	37.0$_{+19.0}$	25.0$_{+16.4}$	22.7$_{+18.0}$	36.2$_{+19.1}$	28.1$_{+3.9}$	15.7$_{+4.9}$	15.6$_{+14.8}$	19.5$_{+3.1}$	26.6$_{-3.1}$	25.2$_{+10.7}$
Text native	Mem0	41.1$_{+23.1}$	25.8$_{+17.2}$	15.6$_{+10.9}$	29.1$_{+12.0}$	26.6$_{+2.4}$	12.9$_{+2.1}$	16.4$_{+15.6}$	14.1$_{-2.3}$	24.2$_{-5.5}$	22.9$_{+8.4}$
MM native	Zep / Graphiti	23.4$_{+5.4}$	13.3$_{+4.7}$	6.2$_{+1.5}$	20.1$_{+3.0}$	19.5$_{-4.7}$	9.0$_{-1.8}$	5.5$_{+4.7}$	11.7$_{-4.7}$	23.4$_{-6.3}$	14.7$_{+0.2}$
Doc render	A-Mem	25.5$_{+7.5}$	16.4$_{+7.8}$	5.5$_{+0.8}$	20.6$_{+3.5}$	10.2$_{-14.0}$	12.6$_{+1.8}$	10.2$_{+9.4}$	12.5$_{-3.9}$	30.5$_{+0.8}$	16.0$_{+1.5}$
Omni ASR	Mem0^g	32.4$_{+14.4}$	23.4$_{+14.8}$	12.5$_{+7.8}$	21.7$_{+4.6}$	26.6$_{+2.4}$	14.4$_{+3.6}$	15.6$_{+14.8}$	18.8$_{+2.4}$	25.8$_{-3.9}$	21.2$_{+6.8}$
	G-Memory	35.2$_{+17.2}$	17.2$_{+8.6}$	13.3$_{+8.6}$	34.9$_{+17.8}$	28.1$_{+3.9}$	18.2$_{+7.4}$	14.8$_{+14.0}$	13.3$_{-3.1}$	21.9$_{-7.8}$	21.9$_{+7.4}$
	MemOS	39.6$_{+21.6}$	25.8$_{+17.2}$	20.3$_{+15.6}$	45.8$_{+28.7}$	34.4$_{+10.2}$	19.2$_{+8.4}$	9.4$_{+8.6}$	17.2$_{+0.8}$	25.0$_{-4.7}$	26.3$_{+11.8}$
	HyperMem†	38.0$_{+20.0}$	23.4$_{+14.8}$	17.2$_{+12.5}$	24.0$_{+6.9}$	25.8$_{+1.6}$	16.5$_{+5.7}$	14.8$_{+14.0}$	9.4$_{-7.0}$	22.7$_{-7.0}$	21.3$_{+6.8}$
	V-Mem	43.0$_{+25.0}$	34.4$_{+25.8}$	25.0$_{+20.3}$	43.4$_{+26.3}$	26.6$_{+2.4}$	12.7$_{+1.9}$	21.1$_{+20.3}$	1.6$_{-14.8}$	22.7$_{-7.0}$	25.6$_{+11.1}$
	MAGE	37.0$_{+19.0}$	20.3$_{+11.7}$	14.1$_{+9.4}$	42.5$_{+25.4}$	27.3$_{+3.1}$	17.9$_{+7.1}$	15.6$_{+14.8}$	14.1$_{-2.3}$	22.7$_{-7.0}$	23.5$_{+9.0}$
	Omni-SimpleMem	38.5$_{+20.5}$	18.8$_{+10.2}$	20.3$_{+15.6}$	25.0$_{+7.9}$	16.4$_{-7.8}$	0.8$_{-10.0}$	19.5$_{+18.7}$	1.6$_{-14.8}$	27.3$_{-2.4}$	18.7$_{+4.2}$
	M3-Agent	34.3$_{+16.3}$	20.3$_{+11.7}$	16.4$_{+11.7}$	22.0$_{+4.9}$	21.9$_{-2.3}$	36.9$_{+26.1}$	16.4$_{+15.6}$	8.6$_{-7.8}$	29.7$_{+0.0}$	22.9$_{+8.5}$
	SuperMem	**52.7**$_{+34.7}$	**35.9**$_{+27.3}$	**28.1**$_{+23.4}$	**58.3**$_{+41.2}$	**55.5**$_{+31.3}$	**48.0**$_{+37.2}$	**21.9**$_{+21.1}$	**21.1**$_{+4.7}$	**31.2**$_{+1.5}$	**39.2**$_{+24.7}$
Gemini 2.5 Flash Lite	No memory	21.0	7.0	2.3	22.3	44.5	17.3	3.9	21.1	22.7	18.0
$0.10/$0.40 per 1M	BM25†	44.7$_{+23.7}$	21.9$_{+14.9}$	16.4$_{+14.1}$	25.1$_{+2.8}$	34.4$_{-10.1}$	18.8$_{+1.5}$	17.2$_{+13.3}$	21.1$_{+0.0}$	25.0$_{+2.3}$	25.0$_{+6.9}$
1M ctx · Jul 2025	MemGPT / Letta	40.8$_{+19.8}$	20.3$_{+13.3}$	25.8$_{+23.5}$	43.7$_{+21.4}$	35.9$_{-8.6}$	16.6$_{-0.7}$	13.3$_{+9.4}$	21.9$_{+0.8}$	15.6$_{-7.1}$	26.0$_{+8.0}$
Text native	Mem0	35.8$_{+14.8}$	25.0$_{+18.0}$	14.1$_{+11.8}$	24.7$_{+2.4}$	31.2$_{-13.3}$	10.2$_{-7.1}$	18.0$_{+14.1}$	24.2$_{+3.1}$	21.1$_{-1.6}$	22.7$_{+4.7}$
MM native	Zep / Graphiti	18.7$_{-2.3}$	10.9$_{+3.9}$	9.4$_{+7.1}$	16.3$_{-6.0}$	28.1$_{-16.4}$	7.4$_{-9.9}$	6.2$_{+2.3}$	22.7$_{+1.6}$	18.8$_{-3.9}$	15.4$_{-2.6}$
Doc native	A-Mem	23.3$_{+2.3}$	14.8$_{+7.8}$	11.7$_{+9.4}$	18.7$_{-3.6}$	27.3$_{-17.2}$	12.7$_{-4.6}$	10.9$_{+7.0}$	20.3$_{-0.8}$	21.1$_{-1.6}$	17.9$_{-0.1}$
Omni native	Mem0^g	21.1$_{+0.1}$	20.3$_{+13.3}$	16.4$_{+14.1}$	16.2$_{-6.1}$	28.1$_{-16.4}$	12.7$_{-4.6}$	14.1$_{+10.2}$	18.8$_{-2.3}$	17.2$_{-5.5}$	18.3$_{+0.3}$
	G-Memory	33.8$_{+12.8}$	21.1$_{+14.1}$	21.1$_{+18.8}$	37.1$_{+14.8}$	41.4$_{-3.1}$	19.0$_{+1.7}$	14.1$_{+10.2}$	25.8$_{+4.7}$	15.6$_{-7.1}$	25.4$_{+7.4}$
	MemOS	39.8$_{+18.8}$	21.9$_{+14.9}$	21.1$_{+18.8}$	42.5$_{+20.2}$	40.6$_{-3.9}$	19.1$_{+1.8}$	12.5$_{+8.6}$	14.8$_{-6.3}$	21.9$_{-0.8}$	26.0$_{+8.0}$
	HyperMem†	34.1$_{+13.1}$	24.2$_{+17.2}$	10.9$_{+8.6}$	27.9$_{+5.6}$	37.5$_{-7.0}$	13.9$_{-3.4}$	14.1$_{+10.2}$	21.1$_{+0.0}$	26.6$_{+3.9}$	23.4$_{+5.4}$
	V-Mem	44.5$_{+23.5}$	25.8$_{+18.8}$	31.2$_{+28.9}$	42.0$_{+19.7}$	38.3$_{-6.2}$	17.2$_{-0.1}$	21.9$_{+18.0}$	21.1$_{+0.0}$	28.9$_{+6.2}$	30.1$_{+12.1}$
	MAGE	34.9$_{+13.9}$	21.1$_{+14.1}$	20.3$_{+18.0}$	43.9$_{+21.6}$	40.6$_{-3.9}$	17.5$_{+0.2}$	14.1$_{+10.2}$	25.0$_{+3.9}$	19.5$_{-3.2}$	26.3$_{+8.3}$
	Omni-SimpleMem	40.9$_{+19.9}$	20.3$_{+13.3}$	21.9$_{+19.6}$	32.5$_{+10.2}$	46.1$_{+1.6}$	8.7$_{-8.6}$	22.7$_{+18.8}$	19.5$_{-1.6}$	21.9$_{-0.8}$	26.1$_{+8.0}$
	M3-Agent	34.0$_{+13.0}$	21.1$_{+14.1}$	21.1$_{+18.8}$	31.2$_{+8.9}$	31.2$_{-13.3}$	50.0$_{+32.7}$	17.2$_{+13.3}$	26.6$_{+5.5}$	27.3$_{+4.6}$	28.9$_{+10.8}$
	SuperMem	**46.1**$_{+25.1}$	**33.6**$_{+26.6}$	**32.0**$_{+29.7}$	**49.3**$_{+27.0}$	**59.4**$_{+14.9}$	**57.4**$_{+40.1}$	**24.2**$_{+20.3}$	**27.3**$_{+6.2}$	**29.7**$_{+7.0}$	**39.9**$_{+21.9}$

4 EXPERIMENTS

We ask eight RQs: **RQ1:** Does SuperMem outperform all baselines? **RQ2:** Does each module help? **RQ3–RQ5:** What do the omni-native super-hypergraph, joint cross-modal recall and conservative consolidation contribute? **RQ6:** Is it efficient? **RQ7:** Is it robust? **RQ8:** How does it answer a case?

4.1 EXPERIMENTAL SETUP

Datasets and Evaluation Metrics. Nine public benchmarks form three blocks by evidence modality (Appendix D): **(A) Text**: LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2025) and MemoryAgentBench (Hu et al., 2026b); **(B) MM**: Mem-Gallery (Bei et al., 2026), MemEye (Guo et al., 2026) and LongDocURL (Deng et al., 2025); **(C) Omni**: M3-Bench-robot (Long et al., 2025), OmniVideoBench (Li et al., 2026a) and LVOmniBench (Tao et al., 2026); metrics are F1 or accuracy.

Baselines and Ablations. Besides No memory, we compare 13 memory baselines by store modality (Appendix E): **(a) text-adapted**: BM25, MemGPT/Letta, Mem0, Zep/Graphiti, A-Mem, Mem0^g, G-Memory, MemOS and HyperMem; **(b) vision-native**: V-Mem and MAGE; **(c) native all-modal**: Omni-SimpleMem and M3-Agent. The three ablations each switch off one module (Appendix E.7).

Table 3: **Ablation of SuperMem's three modules** (`gemini-2.5-flash-lite`). Accuracy is the mean of three draws per block of Table 2; subscripts: change from the full model; ≡: unchanged.

	Accuracy by block			Retrieval		8× writes	
Variant	Text	MM	Omni	R@5	Own@1	ΔR@5	Store
SuperMem (full)	37.2	55.8	23.2	86.6	78.4	±0.0	1.0×
w/o omni-native nodes	≡	$34.2_{-21.7}$	$22.8_{-0.4}$	$82.1_{-4.5}$	$71.7_{-6.6}$	±0.0	1.0×
w/o joint cross-modal recall	≡	$50.5_{-5.3}$	$22.8_{-0.4}$	$81.4_{-5.1}$	$71.2_{-7.2}$	±0.0	1.0×
w/o conservative consolidation	≡	≡	≡	≡	≡	−26.4	8.0×

Implementation Details. All rows share one harness, the encoder ϕ (`gemini-embedding-2`, d=3072) and three frozen backbones f at temperature 0: `gpt-oss-20b` (text), `qwen3.7-flash` (+image, video) and `gemini-2.5-flash-lite` (+audio, file). SuperMem keeps one frozen configuration (Table 9). RQ2–RQ8 use the Gemini backbone and add R@5, the share of gold evidence units in the top five, and Own@1, whether the top unit comes from the question's episode.

4.2 Main Results (RQ1)

As shown in Table 2 and Figure 4, SuperMem has the best average on every backbone, 33.3, 39.2 and 39.9 against 28.2, 26.3 and 30.1 for the best baseline, and over backbones the best row on all nine benchmarks. **The margin is widest where answers lie in media**: +21.1 on MemEye with Qwen, and it leads all nine Omni cells. **Structure and native modality are complementary**: the text hypergraph HyperMem averages 22.6 and the flat vision-native V-Mem 28.0, while SuperMem, combining both, reaches 37.5.

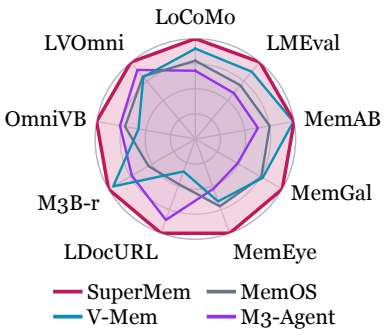


Figure 4: **Main results** averaged over backbones; each axis is scaled to its best row.

4.3 Ablation Study (RQ2)

Table 3 switches off one module at a time. **Omni-native nodes carry the MM block**: replacing every media node by its rendering $r(m)$ costs 21.7 points on MM, 4.5 R@5 and 6.6 Own@1, while the text block, which holds no media, is unchanged. **Joint cross-modal recall reaches the stored media**: scoring edges by text alone costs 5.3 points on MM and the largest retrieval drops (5.1 R@5, 7.2 Own@1). **Conservative consolidation is free and prevents crowding**: it changes no recall on the nine benchmarks, while an append-only store written eight times grows 8× and loses 26.4 R@5.

4.4 Analysis of the Omni-native Super-Hypergraph (RQ3)

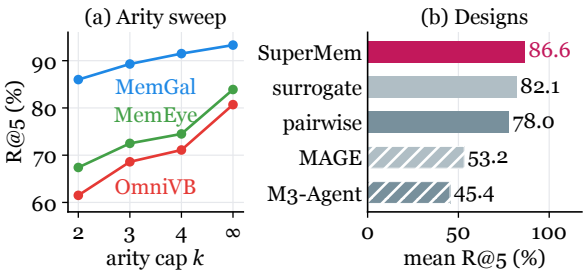


Figure 5: **Structure.** (a) R@5 by arity cap k (k=2: pairwise). (b) Mean R@5 of each design.

As shown in Figure 5, we test the superset of Proposition 1. **Higher arity recovers more evidence**: R@5 never falls as the arity cap k grows, and removing the cap gains +7.3, +16.5 and +19.2 over pairwise k=2 on Mem-Gallery, MemEye and OmniVideoBench (Fig. 5a), since native units are high-order: a clip binds 5.5–6.5 nodes on average and a Mem-Gallery session 9.6, up to 21. **The superset beats its special cases**: SuperMem reaches 86.6 mean R@5, against 82.1 and 78.0 for its surrogate and pairwise restrictions, and 53.2 and 45.4 for MAGE and M3-Agent, which implement them (Fig. 5b).

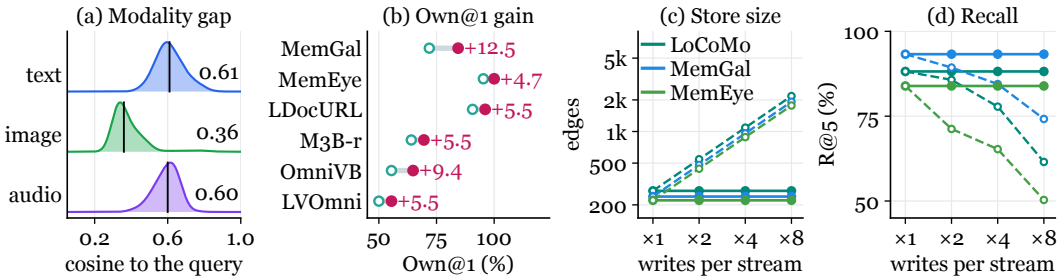


Figure 6: **Recall and consolidation.** (a) Cosine of text probes to stored nodes. (b) Own@1, text-only (open) vs. joint (filled). (c) Store size and (d) R@5 under rewrites; dashed: append-only.

4.5 Joint Cross-Modal Recall across the Modality Gap (RQ4)

As shown in Figure 6a,b, we compare joint recall with text-only scoring on the six media sets. **The modality gap is modality-specific**: a text probe meets image nodes at a median cosine of 0.36, against 0.61 for text and 0.60 for audio, so Eq. 5 credits each modality only above its own median (Prop. 2). **Joint recall finds the right episode and rescues evidence**: it raises Own@1 on all six sets, by up to 12.5 points, and brings 102 gold units into the top five while displacing only 9.

4.6 Conservative Consolidation under Repetition (RQ5)

As shown in Figure 6c,d, we rewrite the streams of LoCoMo, Mem-Gallery and MemEye up to 8×. **The store stays bounded**: every repeat merges into its original edge, so SuperMem keeps 272 edges on LoCoMo however often the stream arrives, while an append-only store grows to 2,176. **Recall is preserved exactly**: R@5 does not move under SuperMem, whereas append-only loses up to 33.6 points as copies of the best units crowd the top five.

4.7 Efficiency: Accuracy at Matched Context (RQ6)

As shown in Figure 7, we cap SuperMem's note budget B and compare accuracy at matched context. **SuperMem leads at every budget.** At B=2k (2,114 recalled characters) it averages 31.9, above every baseline of Table 2, including V-Mem (30.1 on 5,480), and rises steadily to 36.7 at B=8k and 39.9 at the frozen budget (Fig. 7a). **Its context is denser in evidence.** Its first 500 characters hold 58.5% of the gold evidence, more than any baseline holds within 32,000 (Fig. 7b).

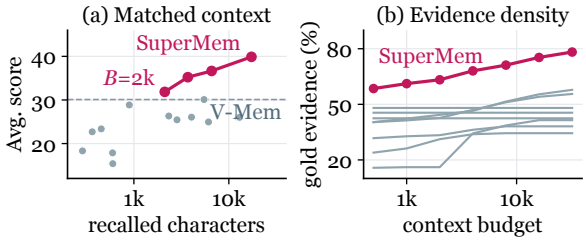


Figure 7: **Efficiency.** (a) Avg. score vs. recalled context; grey: baselines. (b) Gold evidence by budget.

4.8 Robustness and Sensitivity (RQ7)

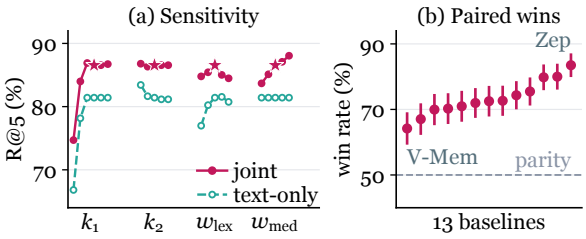
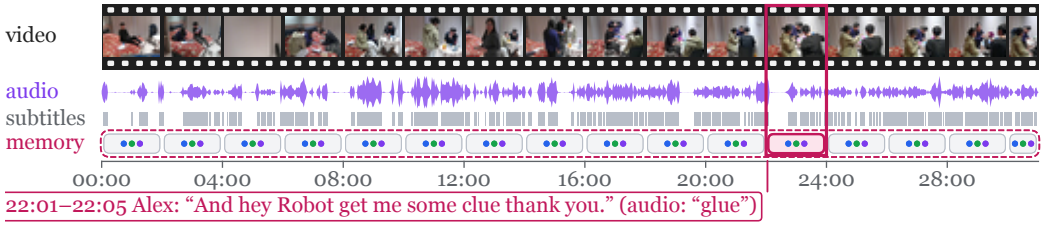


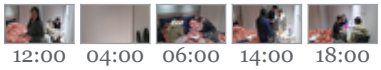
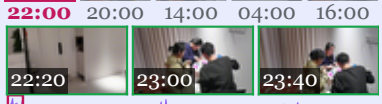
Figure 8: **Robustness.** (a) R@5 under sweeps; stars: frozen. (b) Paired win rates, 95% intervals.

As shown in Figure 8, we sweep four recall hyperparameters and run paired tests. **The frozen values sit on plateaus.** Mean R@5 is flat once $k_1 \geq 6$ and across k_2, peaks at w_{lex}=0.2 and rises with w_{med}, and joint recall beats text-only scoring at every setting (Fig. 8a). **The wins are significant.** On items where they disagree, SuperMem beats each of the 13 baselines on 64–83% of them (sign test, $p < 10^{-8}$; Fig. 8b) and wins or ties in at least 43 of 58 task categories.

Table 4: **Case study** on M3-Bench-robot (`gemini-2.5-flash-lite`). Clips, quotes and answers are verbatim from the run records; pink: the clip holding the request; M3B-r: score in Table 2.

Question (`bedroom_07`, a 31-minute recording written as 16 two-minute clips of frames, audio and subtitles): *On which side of Amanda was Alex sitting when he told the robot to get the glue?* **Gold:** right.



Method	Recalled units (rank order, clip start)	Evidence reaching $f \rightarrow$ answer
MemOS T M3B-r 12.5	*text* *text* *text* *text* *text* other 04:00 24:00 16:00 other	text only, no file → "The provided evidence does not contain an instance . . . " ✗
HyperMem T M3B-r 14.1	*fact* *fact* *fact* *fact* *fact* five facts, none tied to a clip or a file	"Alex asked Robot to find some glue." → "Information not available." ✗
V-Mem M M3B-r 21.9	 12:00 04:00 06:00 14:00 18:00	12 files, all of clips without the request → "Not mentioned" ✗
MAGE M M3B-r 14.1	*text* **22:00**	the right clip, but as text, no file → "Cannot be determined . . . " ✗
Omni-SimpleMem O M3B-r 22.7	 38:00 12:00 12:00 08:00 04:00	12 files, two of other videos → "The provided notes do not contain . . . " ✗
M3-Agent O M3B-r 17.2	 06:00	4 files of one clip, 06:00–08:00 → "Not applicable" ✗
SuperMem O M3B-r **24.2**	 **22:00** 20:00 14:00 04:00 16:00	the request's clip first: s=0.76 (text 0.73, image 0.34, audio 0.63)
clip 22:00–24:00 → s = **0.76**	 22:20 23:00 23:40 audio 22:00–24:00	the whole hyperedge reaches f, note, frames and audio together: "The image shows Alex on the right, and Amanda on the left." → **"Right"** ✓

4.9 Case Study (RQ8)

As shown in Table 4, we trace a question over a 31-minute recording whose answer needs two modalities at one moment: the spoken request and the seating visible in its frames. **Baselines lose either the moment or its media.** MemOS and HyperMem recall text alone, HyperMem even the request itself but with no time or frame, and MAGE finds the right clip only as text; V-Mem, Omni-SimpleMem and M3-Agent deliver frames, but of other clips or recordings, and all six report that the evidence lacks it. **SuperMem recalls the moment whole.** Its joint score ranks the request's clip first, and since the hyperedge is returned whole, its frames travel with the request and the backbone reads the seating: "right".

5 Conclusion

We presented SuperMem, an omni-native super-hypergraph memory for agents recalling evidence across text, images, documents, audio and video. It stores each unit with its native media as a first-level hyperedge and nests an episode's units in a second-level hyperedge, recalls whole units by a joint cross-modal score, and consolidates conservatively, merging repeats without deleting. On nine text, multimodal and omni-modal benchmarks with three frozen backbones, SuperMem attains the best average against 13 memory baselines, and analyses confirm each module, denser evidence at matched context and robust recall. Extending it to agents acting on memory is a natural next step.

AI USE STATEMENT

We used generative AI tools only to polish the writing and to assist with code implementation, while the research questions, method design, experiments, analyses and conclusions are the authors' own. All AI-assisted text and code were reviewed, verified and tested by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work uses only publicly released benchmarks as distributed by their authors and involves no human subjects, crowdsourcing or newly collected personal data. We have read and adhere to the ICLR Code of Ethics, and declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The anonymous repository linked in the abstract releases the SuperMem source code with the frozen configuration used in every run (Table 9), and Appendices A and C give all prompts and the full algorithm with its complexity. Appendix B states the assumptions and complete proofs of Propositions 1–3, and Appendices D, E and F detail the benchmarks, evaluation protocol, baseline configurations and per-experiment settings.

REFERENCES

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 Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

Weihao Bo, Shan Zhang, Yanpeng Sun, Jingjing Wu, Qunyi Xie, Xiao Tan, Kunbin Chen, Wei He, Xiaofan Li, Na Zhao, Jingdong Wang, and Zechao Li. Vilomem: Agentic learner with grow-and-refine multimodal semantic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5476–5486, June 2026.

Ruike Cao, Fanyu Zhao, Fugen Yao, Liang Dong, Jian Xu, Guanjun Jiang, Yifei Zhao, Han Zhang, and Li Xiao. Memcalib: Benchmarking and optimizing memory use in llm agents, 2026a. URL https://arxiv.org/abs/2609.24259.

Zijie Cao, Xijun Qu, Zhicheng Gu, Xiaoshu Chen, Duanyang Yuan, Yanning Hou, Sihang Zhou, Jianxing Gong, Jian Huang, and Yang Mei. Autoviewmem: Self-configuring orthogonal views for conversational long-term memory, 2026b. URL https://arxiv.org/abs/2609.21940.

Chunliang Chen, Ming Guan, Xiao Lin, Jiaxu Li, Luxi Lin, Qiyi Wang, Xiangyu Chen, Jixiang Luo, Changzhi Sun, Dell Zhang, and Xuelong Li. Telemem: Building long-term and multimodal memory for agentic ai, 2026a. URL https://arxiv.org/abs/2601.06037.

Han Chen, Zining Zhang, Wenqi Pei, Bingsheng He, Ming Wu, Jason Zeng, Michael Heinrich, Wei Wu, and Hongbao Zhang. Memforest: An efficient agent memory system with hierarchical temporal indexing, 2026b. URL https://arxiv.org/abs/2605.23986.

Yijun Chen, Boyi Xiao, Yixian Zhao, Haoting Xia, Buqiang Xu, Jizhan Fang, Yanya Li, Yaqi Zheng, Xuehai Wang, Zirui Xue, Liuxin Zhang, Hui Li, and Ningyu Zhang. Lightmem-ego: Your ai memory for everyday life, 2026c. URL https://arxiv.org/abs/2607.11487.

Yijun Chen, Yaqi Zheng, Yanya Li, Boyi Xiao, Buqiang Xu, Shuofei Qiao, Jizhan Fang, Xinle Deng, Yunzhi Yao, Xuehai Wang, Liuxin Zhang, Hui Li, Huajun Chen, and Shumin Deng. EM^{2}Mem: Event-centric multimodal memory for large language models, 2026d. URL https://arxiv.org/abs/2609.00551.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Zeyang Cui, Jiannong Cao, Zhiyuan Wen, Bo Yuan, Junlan Feng, and Shengyuan Chen. Edgemem: Llm-free agent memory construction and retrieval via evidence-preserving multi-anchor hypergraph, 2026. URL https://arxiv.org/abs/2609.05553.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.57. URL https://aclanthology.org/2025.acl-long.57/.

Jiajun Dong, Yutao Hu, Fengrui Fan, Shihan Dou, Yueming Wu, and Deqing Zou. Memarbiter: Decision-time memory arbitration for long-horizon llm agents, 2026. URL https://arxiv.org/abs/2608.02113.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. Lightmem: Lightweight and efficient memory-augmented generation, 2026a. URL https://arxiv.org/abs/2510.18866.

Jizhan Fang, Buqiang Xu, Zhixian Wang, Haoliang Cao, Xinle Deng, Baohua Dong, Hangcheng Zhu, Ruohui Huang, Gang Yu, Ying Wei, Guozhou Zheng, Feiyu Xiong, Haofen Wang, Huajun Chen, and Ningyu Zhang. Rethinking memory as continuously evolving connectivity, 2026b. URL https://arxiv.org/abs/2605.28773.

Yichao Feng, Ran Zhang, Haoran Luo, Zhenghong Lin, Carl Yang, and Anh Tuan Luu. Diachronic hypergraphs for orchestrated multi-agent multimodal memory curation, 2026. URL https://arxiv.org/abs/2608.29678.

Xian Gao, Jinpeng Wang, Jiacheng Ruan, Guangyu Cao, Ting Liu, and Yuzhuo Fu. Memo: Multimodal evidence memory organization for long-horizon llm agents, 2026. URL https://arxiv.org/abs/2609.07471.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, and Ruixiang Tang. Memeye: A visual-centric evaluation framework for multimodal agent memory, 2026. URL https://arxiv.org/abs/2605.15128.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. EverMemOS: A self-organizing memory operating system for structured long-horizon reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45836–45853, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2125. URL https://aclanthology.org/2026.acl-long.2125/.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions, 2026b. URL https://arxiv.org/abs/2507.05257.

Zhanghao Hu, Qinglin Zhu, Runcong Zhao, Di Liang, Hanqi Yan, Yulan He, and Lin Gui. Beyond rag for agent memory: Retrieval by decoupling and aggregation, 2026c. URL https://arxiv.org/abs/2602.02007.

Shuo Ji, Yibo Li, and Bryan Hooi. Memory is reconstructed, not retrieved: Graph memory for llm agents, 2026. URL https://arxiv.org/abs/2606.06036.

Dongming Jiang, Yi Li, and Bingzhe Li. Jev-mem: System-one-controlled agentic memory for efficient ai agents, 2026. URL https://arxiv.org/abs/2609.23986.

Dingyi Kang, Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. V-mem: Modality-routed retrieval for long-term multimodal agentic memory, 2026. URL https://arxiv.org/abs/2608.01543.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In Greg Durrett and Ping Jian (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 275–285, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-392-0. doi: 10.18653/v1/2026.acl-demo.27. URL https://aclanthology.org/2026.acl-demo.27/.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms, 2026a. URL https://arxiv.org/abs/2510.10689.

Geng Li, Yuhao Wang, Dong Li, Jianye Hao, and Yuxin Peng. Graphmemix: Query-aware evidence forests for long-term multimodal agent memory, 2026b. URL https://arxiv.org/abs/2608.26983.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Chen Tang, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhen Tao, Huayi Lai, Hao Wu, Bo Tang, Zhengren Wang, Zhaoxin Fan, Ningyu Zhang, Linfeng Zhang, Junchi Yan, Mingchuan Yang, Tong Xu, Wei Xu, Huajun Chen, Haofen Wang, Hongkang Yang, Wentao Zhang, Zhi-Qin John Xu, Siheng Chen, and Feiyu Xiong. Memos: A memory os for ai system, 2025. URL https://arxiv.org/abs/2507.03724.

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 Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11601–11617, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.533. URL https://aclanthology.org/2026.acl-long.533/.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625. Curran Associates, Inc., 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf.

Yidan Lin, Kaixiang Wang, Jiong Lou, and Jie Li. Stop when memory suffices: Evidence-conditioned progressive execution for llm agents, 2026. URL https://arxiv.org/abs/2608.01285.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, Mingyu Ding, and Huaxiu Yao. Omni-simplemem: Autoresearch-guided discovery of lifelong multimodal agent memory, 2026a. URL https://arxiv.org/abs/2604.01007.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for llm agents, 2026b. URL https://arxiv.org/abs/2601.02553.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, Yi Yu, Shuyue Hu, Botian Shi, and Ding Wang. Memverse: Multimodal memory for lifelong learning agents, 2026c. URL https://arxiv.org/abs/2512.03627.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory, 2025. URL https://arxiv.org/abs/2508.09736.

Yihao Lu, Wanru Cheng, Zeyu Zhang, and Hao Tang. Mma: Multimodal memory agent, 2026. URL https://arxiv.org/abs/2602.16493.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 152206–152234. Curran Associates, Inc., 2025. doi: 10.52202/085713-5089. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/df55ee6e59f8ac4a625219e11fe9ddba-Paper-Conference.pdf.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

Theo Rusu, Sourena Khanzadeh, and Manar Alalfi. Selective forgetting: A graph-based memory framework for long-term llm agents, 2026. URL https://arxiv.org/abs/2608.28978.

Florentin Smarandache. Extension of hypergraph to n-superhypergraph and to plithogenic n-superhypergraph, and extension of hyperalgebra to n-ary (classical-/neutro-/anti-) hyperalgebra. *Neutrosophic Sets and Systems*, 33:290–296, 2020.

Yuanyi Song, Yukai Wang, Xinbei Ma, Zhihui Fu, Jianghao Lin, Weiwen Liu, Jun Wang, Huarong Deng, Yong Yu, and Weinan Zhang. Retrieval-driven memory reconsolidation for long-term llm agents, 2026. URL https://arxiv.org/abs/2609.16053.

Jingyu Sun, Yan Lin, Yuyang Xue, Yifan Wang, Zhengtao Yao, Rui Qian, Zefeng Xu, Jiachen Li, Xianyang Liu, Jiancheng Pan, Jingyuan Sun, Syed Murtuza Baker, and Hongpeng Zhou. Pmmc: Prospective multimodal memory compilation for long-term lvlm agents, 2026. URL https://arxiv.org/abs/2608.00962.

Zheling Tan, Jin Gao, and Dequan Wang. Cable: Extending the reach of memory retrieval via complementary antecedent-based linking and expansion, 2026. URL https://arxiv.org/abs/2608.17911.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, Weiqiang Wang, Jian Liu, Can Qin, Yulun Zhang, Ming-Hsuan Yang, and Huan Wang. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms, 2026. URL https://arxiv.org/abs/2603.19217.

Zhoujin Tian, Hao Zhang, Yao Tian, Cheng Chen, Yakun Li, Lei Zhang, and Xiaofang Zhou. Beyond retrieval: Analytic memory for multimodal agents, 2026. URL https://arxiv.org/abs/2607.29440.

Fengrong Wan, Chengcan Wu, and Ningtao Lyu. Sodamem: Evidence-grounded temporal graph memory for llm agents, 2026. URL https://arxiv.org/abs/2608.08055.

Junxi Wang, Te Sun, Jiayi Zhu, Chen Zhang, Siyuan Li, Xuyang Liu, Zichen Wen, Xiaobing Tu, Jinkui Ren, Xiantao Zhang, Ziqi Yuan, and Linfeng Zhang. Memforest: Efficient agent memory management via eventtree partitioning and progressive merging, 2026a. URL https://arxiv.org/abs/2609.08273.

Qiuchen Wang, Shihang Wang, Yu Zeng, Qiang Zhang, Fanrui Zhang, Zhuoning Guo, Bosi Zhang, Wenxuan Huang, Lin Chen, Zehui Chen, Pengjun Xie, and Ruixue Ding. Vimrag: Navigating massive visual context in retrieval-augmented generation via multimodal memory graph, 2026b. URL https://arxiv.org/abs/2602.12735.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents, 2025. URL https://arxiv.org/abs/2507.07957.

Ziming Wang. Toki: A bitemporal operator algebra for contradiction resolution in llm-agent persistent memory, 2026. URL https://arxiv.org/abs/2606.06240.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu, and Hongwei Wang. GAM: Hierarchical graph-based agentic memory for LLM agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34647–34664, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1600. URL https://aclanthology.org/2026.acl-long.1600/.

Zhifei Xie, Jiaqi Lang, Ze An, Yifan Zhao, Dongchao Yang, Kai Li, Ziyang Ma, Mingbao Lin, Chunyan Miao, and Shuicheng Yan. Voicemem: Streaming dual-brain memory for real-time interaction, 2026. URL https://arxiv.org/abs/2608.26005.

Ruiyao Xu, Tiankai Yang, and Wei-Chieh Huang. Hyperskill: Self-evolving llm agents via hypergraph-structured skill memory, 2026. URL https://arxiv.org/abs/2608.16114.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

Meriem Yacoubi, Pia Schmidt, Nenad Petrovic, Ahmed Frikha, Martin Kirchhoff, and Alois Knoll. Memorylace: Memory lifecycle-aware consolidation and evidence retrieval, 2026. URL https://arxiv.org/abs/2609.03201.

Chang Yang, Chuang Zhou, Yilin Xiao, Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, Zhishang Xiang, Shengyuan Chen, Huachi Zhou, Qinggang Zhang, Ninghao Liu, Jinsong Su, Xinrun Wang, Yi Chang, and Xiao Huang. Graph-based agent memory: Taxonomy, techniques, and applications, 2026a. URL https://arxiv.org/abs/2602.05665.

Kairui Yang, Minghao An, Xunkai Li, Ziheng Yi, Zekai Chen, Guangyuan He, and Rong-Hua Li. Mace: Memory-agent co-evolution with adaptive memory graphs for multi-agent systems, 2026b. URL https://arxiv.org/abs/2609.21533.

Ke Yang, Zixi Chen, Xuan He, Jize Jiang, Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, and ChengXiang Zhai. Plugmem: A task-agnostic plugin memory module for llm agents, 2026c. URL https://arxiv.org/abs/2603.03296.

Qing Yang, Pengcheng Huang, Xinze Li, Zhenghao Liu, Yukun Yan, Yu Gu, Ge Yu, Gang Li, and Maosong Sun. Memorycard: Topic-aware multi-modal clue compression for long-video question answering, 2026d. URL https://arxiv.org/abs/2606.05917.

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 (CVPR)*, pp. 25599–25609, June 2026.

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo, and Yafeng Deng. HyperMem: Hypergraph memory for long-term conversations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35237–35254, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1627. URL https://aclanthology.org/2026.acl-long.1627/.

Xiawei Yue, Boran Wang, Xiaoqing Zhang, Shuxin Zheng, and Ziwei Zhang. Hierarchical graph memory for llm agents with path-level localization and rewrite, 2026b. URL https://arxiv.org/abs/2608.05095.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 12988–13018. Curran Associates, Inc., 2025. doi: 10.52202/085713-0439. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/136a45cd9b841bf785625709a19c6508-Paper-Conference.pdf.

Huixin Zhang, Shao-Jun Xia, Di Wang, Liangxi Liu, Hainan Xiong, and Zihao Wang. Collaborative memory for multi-agent vlm systems, 2026. URL https://arxiv.org/abs/2609.17921.

Beidi Zhao, Yaoqi Chen, Yuru Feng, Menghao Li, Qianxi Zhang, Baotong Lu, Jianan Lu, Zhirui Wang, Xinjiang Wang, Shusen Xu, Zengzhong Li, Xiaoxiao Li, and Qi Chen. Mesa:task-adaptive multi-structure evidence selection for long-horizon agent memory, 2026. URL https://arxiv.org/abs/2608.10108.

Chulun Zhou, Chunkang Zhang, Guoxin Yu, Fandong Meng, Jie Zhou, Wai Lam, and Mo Yu. Hgmem: Hypergraph-based working memory to improve multi-step rag for long-context complex relational modeling, 2026a. URL https://arxiv.org/abs/2512.23959.

Heng Zhou, Lian Zhang, Yutao Fan, Tiancheng He, Siki Chen, Hejia Geng, Philip Torr, and Zhenfei Yin. Latticemind: A conflict-aware memory primitive for multi-agent systems, 2026b. URL https://arxiv.org/abs/2608.08236.

Wenqi Zhou, Zhuorui Yu, Kaiao Wen, Hao Zheng, Xinyi Zheng, Peiran Wu, Enmin Zhou, Chi-Hao Wu, and Junxiao Shen. To memories and beyond: From remembering to knowing you across long-term multimodal personal archives, 2026c. URL https://arxiv.org/abs/2609.19167.

APPENDIX

A PROMPTS USED IN SUPERMEM

SuperMem itself runs without prompts: writing, recall and consolidation follow Eqs. 2–11. Prompts enter only in the shared rendering $r(m)$ of each file (§A.1) and in the frozen backbone f of Eq. 1 (§A.2), both identical for every row of Table 2. All are verbatim, { } marks a field filled per call, and every call runs at temperature 0.

A.1 SHARED MEDIA RENDERING PROMPT

Each file m is rendered once by `gemini-2.5-flash-lite`, whatever the backbone, with the prompt of its modality (Figure 9). The reply, at most 160 tokens and tagged with its modality (e.g. `[image]`), is $r(m)$, cached per file so that every row, backbone and question reads the same text. In SuperMem, $r(m)$ joins the unit text $\tilde{c}_t$ (Eq. 2), reaching the unit node and the lexical term of Eq. 5, while media nodes embed the file itself; a note carries $r(m)$ in place of a file f cannot read (Eq. 8).

-Image-
Describe this image so it can be retrieved later by a text query: what it shows, the entities in it, and any text or numbers it contains. Two to three sentences.

-Audio-
Transcribe the speech in this audio verbatim, then name the non-speech sounds, the speakers and the setting. If there is no speech, describe the sounds. Two to three sentences.

-Video-
Describe this video so it can be retrieved later by a text query: what happens and in what order, who or what appears, any on-screen text, and what is said. Two to three sentences.

Figure 9: Prompts of the shared rendering $r(m)$ in Eq. 2; the file m precedes the prompt in one user turn. Figure 11 shows renderings as the lines tagged `[image]` and `[audio]`.

A.2 ANSWERING PROMPTS OF THE BACKBONE

The backbone f is a planner and a solver on one model (Figure 10): the planner turns the recalled notes and the question into at most three steps (200 tokens), and the solver answers from the delivered files $\mathcal{F}_q$, the notes, the plan and the question (1,600 tokens). A note is one recalled edge e: a header with its stream, the text part $s_q^{\text{txt}}(e)$, the similarities $s_q^{\mu}(e)$ of Eq. 5 and its files, then its text $\tilde{c}_e$, whole for the top $n_{\text{full}} = 5$ notes and an $\ell_{\text{w}} = 1{,}200$-character window otherwise.

-Planner- system You coordinate a small team solving one question. Given the question and any recalled notes, write at most three short steps the solver should follow. No preamble. user {notes}↩↩Question: {question}
-Solver- system You answer the question using the attached evidence and the plan. Think if you need to, but keep the reasoning to about ten sentences at most, then end your reply with a final line of exactly this form:↩ANSWER: <your answer> user {files} {notes}↩↩Plan:↩{plan}↩↩{question block}
-Question block- Question: {question}↩↩{answer instruction}↩Think if you need to, then end your reply with a final line of exactly this form:↩ANSWER: <your answer> *[after, where they apply, the rendered evidence, item note and options, each ending ↩↩]*
-Answer instruction- *open:* Answer as briefly as possible. *options, letter gold:* Answer with the option letter only. *options, text gold:* Answer with the full text of the correct option.
-Options- Options:↩{options}↩↩You must pick one of the options {above | listed with the question below}, even if you are unsure. Do not answer with anything that is not one of them. *[the list is omitted when the question carries it]*
-Item note- *image id, Mem-Gallery:* Answer with the image id exactly as it appears in the conversation, e.g. D4:IMG_006. If several images qualify, list their ids. *label, MemoryAgentBench:* The recorded stream assigns each question a numeric label. Answer with the label number for this question and nothing else. *stream fact, MemoryAgentBench:* Answer from the recorded stream. It states facts that may contradict what you already know; where they differ, the stream is correct.
-Rendered evidence- *files f cannot read:* Rendered evidence (media this model cannot read, described once by a shared pass):↩{$r(m)$ of each such file}
-Assistant- *unusable reply:* system The solver failed to produce a usable answer. Using the attached evidence, give the answer yourself. Do not restate the question. End with a final line of exactly this form:↩ANSWER: <your answer> user {files} {notes}↩↩Question: {question}↩↩The solver's unusable reply was:↩{reply, first 400 characters}
-Extractor- *no parseable answer:* system Read the passage and output ONLY the final answer it arrives at -- a single option letter if it chose one, otherwise the shortest exact phrase or value. No explanation, no punctuation around it. user {reply, last 2,500 characters}

Figure 10: Prompts of the backbone f in Eq. 1, shared by every row of Table 2; ↩ marks a line break and grey our annotations. The last two act only on an unusable reply (App. D.2).

system You coordinate a small team solving one question. Given the question and any recalled notes, write at most three short steps the solver should follow. No preamble.
user Recalled notes:
- [recalled from stream bedroom_07; matched by text 0.73, image 0.34, audio 0.63; files: f00034, f00035, f00036, a00011]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 11 (22:00-24:00)
Spoken: Jinx:Round eyes yes okay. Alex:And hey Robot get me some clue thank you. Robot:Ok. Jinx:We need some eyes...small eyes or big eyes? Alex:Small eyes okay. Amanda:I want big eyes okay. Amanda:What about two small eyes and one big eyes okay. Alex:Okay. Amanda:I will do it by myself. Jinx:Okay. Jinx:Just slow down and pay attention to it don't hurt yourself. Jinx:Don't hurt your fingers. Jinx:Wow one eyes. Alex:Wow. Amanda:It's so small. Jinx:Yeah I think that's right I like the small. Jinx:Slow down. Jinx:Wow oh this big eyes... Jinx:The edge is a little sharp. Amanda:Really sharp okay. Jinx:Where's the glue? Alex:Oh. Jinx:Thank you. Alex:It's really funny right? Alex:Yeah I like it... Jinx:So small. Alex:Wow it's so cute... Amanda:Like this.
[image] The image shows a modern entryway with white walls and marble stairs. A black chair is visible in the foreground, and a pair of shoes rests on a mat near a white door. The scene is brightly lit, creating a clean and minimalist aesthetic.
[image] Three people are gathered around a small table, appearing to be engaged in a craft or activity. The person in the center, wearing a blue sweater, is focused on something on the table, while the person to their left, in a tan jacket, holds a small object. A person in a black sweatshirt is seated on the right, also looking towards the table.
[image] In this image, three people are gathered around a table, seemingly engaged in a craft or activity. The scene is indoors, with curtains in the background and a wall with electrical outlets. One person is wearing a tan jacket, another a dark sweater, and the third a dark top.
[audio] Okay, and uh, hey robot, can you find some glue? Thank you. Okay.
Non-speech sounds:
The audio contains the sound of someone speaking, interspersed with the sound of a robot's voice. There are also sounds of paper rustling and a slight whirring noise, possibly from the robot.
Speakers:
There are two speakers: a human and a robot.
Setting:
The setting appears to be a workshop or a craft area, given the mention of glue and the sounds of paper.
- [recalled from stream bedroom_07; matched by text 0.66, image 0.35, audio 0.57; files: f00031, f00032, f00033, a00010]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 10 (20:00-22:00)
Spoken: Jinx:No I think this is so difficult. Alex:And just fold it. Jinx:Fold it about once or [...]
- [recalled from stream bedroom_07; matched by text 0.65, image 0.33, audio 0.60; files: f00022, f00023, f00024, a00007]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 07 (14:00-16:00)
Spoken: Jinx:You're talking about the film right? Amanda:yes. Amanda:So you wanna teach us to do [...]
- [recalled from stream bedroom_07; matched by text 0.66, image 0.33, audio 0.61; files: f00007, f00008, f00009, a00002]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 02 (04:00-06:00)
Spoken: Alex:I look this way ... Alex:Very interesting... Jinx:It's interesting... Jinx:Wow! [...]
- [recalled from stream bedroom_07; matched by text 0.64, image 0.35, audio 0.58; files: f00025, f00026, f00027, a00008]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 08 (16:00-18:00)
Spoken: Amanda:What should we do now? Alex:But we don't have the scissors yeah. Alex:So Jinx can [...]
- [recalled from stream bedroom_07; matched by text 0.63, image 0.34, audio 0.63; files: f00013, f00014, f00015, a00004]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 04 (08:00-10:00)
[...] to a bed, looking down at a small table. A young man is seated at the table, engrossed in a tablet, [...]
- [recalled from stream bedroom_07; matched by text 0.64, image 0.34, audio 0.61; files: f00019, f00020, f00021, a00006]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 06 (12:00-14:00)
[...] Alex:Ah oh but it's a little bit boring. Amanda:Boring yeah not horrible at all. Alex:Yes... [...]
- [recalled from stream bedroom_07; matched by text 0.61, image 0.32, audio 0.68; files: f00010, f00011, f00012, a00003]
[hyperedge over text+image+audio] Recorded stream bedroom_07 clip 03 (06:00-08:00)
[...] and white pants, while the person in the middle is wearing a blue jacket. The person on the right [...]
Question: On which side of Amanda was Alex sitting when he told the robot to get the glue?
assistant (planner) 1. Identify the specific instance where Alex asks the robot for glue.
2. Locate the description of the seating arrangement or relative positions of Alex and Amanda at that moment.
3. Determine Alex's position relative to Amanda.
system You answer the question using the attached evidence and the plan. Think if you need to, but keep the reasoning to about ten sentences at most, then end your reply with a final line of exactly this form:
ANSWER: <your answer>
user ⟨**f00034.jpg**⟩ ⟨**f00035.jpg**⟩ ⟨**f00036.jpg**⟩ ⟨**a00011.mp3**⟩ ⟨**f00031.jpg**⟩ ⟨**f00032.jpg**⟩ ⟨**f00033.jpg**⟩ ⟨**a00010.mp3**⟩ ⟨**f00022.jpg**⟩ ⟨**f00023.jpg**⟩ ⟨**f00024.jpg**⟩ ⟨**a00007.mp3**⟩
Recalled notes:
[the eight notes above, unchanged]
Plan:
1. Identify the specific instance where Alex asks the robot for glue.
2. Locate the description of the seating arrangement or relative positions of Alex and Amanda at that moment.
3. Determine Alex's position relative to Amanda.
Question: On which side of Amanda was Alex sitting when he told the robot to get the glue?
Answer as briefly as possible.
Think if you need to, then end your reply with a final line of exactly this form:
ANSWER: <your answer>
assistant (solver) Alex asked the robot to get glue while he was sitting on the right side of Amanda. The image shows Alex on the right, and Amanda on the left, both looking at the table.
ANSWER: Right

Figure 11: The complete prompt of the case in Table 4: the planner call, then the solver call with the twelve files of the top $n_e = 3$ notes (M3-Bench-robot, `gemini-2.5-flash-lite`). All eight notes come from the recording asked about, the top one is the clip in which Alex makes the request, and neither safeguard of Figure 10 fires. Text is verbatim up to blank lines; grey marks our annotations, and [...] note text elided for space.

B THEORETICAL PROOF

This appendix proves the three propositions of §3 in the notation defined there. Every statement concerns the construction of SuperMem itself, so each proof is exact and holds for any encoder ϕ, backbone f and benchmark $\mathcal{D}$.

B.1 PROOF OF PROPOSITION 1

Proposition 1. *Text-surrogate, pairwise and flat memories are all special cases of SuperMem.*

Proof. Call a memory design a special case of SuperMem if a restriction of the store $\mathcal{H} = (\mathcal{V}, \mathcal{E}^{(1)}, \mathcal{E}^{(2)})$ of Eq. (3), or of its hyperparameters, reproduces the design's store while Eqs. (5)–(8) keep their form. We exhibit such a restriction for each design.

Text surrogate. Replace every unit's node set by its text part, $V_t \mapsto V_t^{\mathsf{txt}} = \{v_t^{\circ}\} \cup \{v_{t,j}\}_{j=1}^{J_t}$, and deliver no raw files. Because $\tilde{c}_t = c_t \oplus r(\mathcal{M}_t)$ already contains the rendering of every file, each edge stores exactly the unit's text together with a textual surrogate of its media, which is the record a surrogate memory keeps. In Eq. (5) $\Sigma_e = \varnothing$ for every edge, so the media term vanishes and

$$s(e) = s_q^{\mathsf{txt}}(e) = (1 - \lambda_q) \max_{v \in e^{\mathsf{txt}}} \cos(z_q, z_v) + \lambda_q \, b_q(e), \qquad \mathcal{F}_q = \varnothing, \tag{12}$$

which is dense-plus-lexical retrieval over captioned units.

Pairwise. Replace each e_t by the $|\mathcal{M}_t|$ edges $e_t^{(m)} = e_t^{\mathsf{txt}} \cup \{v_m\}$, $m \in \mathcal{M}_t$ (and by e_t^{txt} alone when $\mathcal{M}_t = \varnothing$). Each edge then relates the unit's text to one file, i.e. the unit is decomposed into binary text–medium relations. This is the arity-restricted configuration $|e^{\mathrm{med}}| \le 1$, under which $|\Sigma_e| \le 1$ and the media term of Eq. (5) reduces to the pairwise relevance $w_{\mathrm{med}}[s_q^{\kappa(m)}(e_t^{(m)}) - \tilde{s}_q^{\kappa(m)}]_+$ of the text and one file.

Flat. Take $k_2 \ge |\mathcal{E}^{(2)}|$, so that the super level is skipped. Then $\mathrm{Top}_{k_2}(g) = \mathcal{E}^{(2)}$ and, since every first-level edge belongs to the hyperedge E_S of its stream, $\mathcal{C}_q = \mathcal{L}_q \cap \bigcup_S E_S = \mathcal{L}_q$, so g no longer influences $\mathcal{K}_q$ and recall is a top-k_1 search over first-level edges, which is a flat memory.

Strict containment. The converse inclusions fail: the surrogate holds no vector of any medium, the pairwise memory has no edge containing two media of one unit, and the flat memory has no set of edges that is itself scored, whereas $\mathcal{H}$ has all three. Hence the two-level hyperedge strictly contains each design.

Whole-unit delivery. Containment also has an operational consequence. Let a unit x_t carry $n = |\mathcal{M}_t| \ge 2$ files. In SuperMem, $e_t \in \mathcal{K}_q$ makes all n files candidates for $\mathcal{F}_q$ through a single one of the k_1 slots. In the pairwise memory a file m is a candidate only if its own edge $e_t^{(m)}$ is kept, so delivering the unit whole takes n slots, and it is impossible when $n > k_1$. Under the same FILL caps, SuperMem therefore delivers every unit that the pairwise memory delivers whole while spending fewer slots on it.

In summary, every prior design is a restriction of SuperMem, which, unlike any of them, binds a whole multimodal unit in one edge. □

B.2 PROOF OF PROPOSITION 2

Proposition 2. *The media bonus never lowers a score and is immune to a constant modality gap.*

Proof. By Eq. (5), $s(e) = s_q^{\mathsf{txt}}(e) + \beta_q(e)$, where $s_q^{\mathsf{txt}}(e)$ is the text and lexical part and the media bonus is

$$\beta_q(e) = \frac{w_{\mathrm{med}}}{|\Sigma_e|} \sum_{\mu \in \Sigma_e} \left[s_q^{\mu}(e) - \tilde{s}_q^{\mu} \right]_+ \quad (\beta_q(e) = 0 \text{ if } \Sigma_e = \varnothing). \tag{13}$$

Never lowers. Every summand is non-negative and $w_{\mathrm{med}} \ge 0$, so $\beta_q(e) \ge 0$ and $s(e) \ge s_q^{\mathsf{txt}}(e)$. The time factor $\rho_q(e) \in (0, 1]$ of Eq. (6) and the factor ω_{sup} of Eq. (11) are positive and do not depend on Σ_e, so for two edges e_1, e_2 with $s_q^{\mathsf{txt}}(e_1) = s_q^{\mathsf{txt}}(e_2)$ and equal such factors, of which only

e_1 holds media, $\hat{s}(e_1) \geq \hat{s}(e_2)$: holding a medium can never rank an edge below an otherwise equal text-only edge.

Immune to a constant gap. Model the gap of modality μ as an offset Δ_μ, which may depend on μ and on the query but not on the edge, on every query–node similarity of that modality, $s_q^\mu(e) = \check{s}^\mu(e) - \Delta_\mu$ for all edges e holding μ, where $\check{s}^\mu$ is the gap-free similarity. The median commutes with translation, $\mathrm{med}(\check{s}^\mu - \Delta_\mu) = \mathrm{med}(\check{s}^\mu) - \Delta_\mu$, hence

$$s_q^\mu(e) - \tilde{s}_q^\mu = \big(\check{s}^\mu(e) - \Delta_\mu\big) - \big(\mathrm{med}\,\check{s}^\mu - \Delta_\mu\big) = \check{s}^\mu(e) - \mathrm{med}\,\check{s}^\mu, \tag{14}$$

and $\beta_q(e)$, and with it the ranking of all edges, is the same for every value of Δ_μ.

Selective and bounded. Let n_μ edges hold μ. At most $\lfloor n_\mu/2 \rfloor$ of them lie strictly above their median, so at most half of them receive a positive μ-term, and a single edge holding μ receives none. Since cosine similarities lie in $[-1, 1]$, $0 \leq \beta_q(e) \leq w_{\mathrm{med}} \max_{\mu \in \Sigma_e}(1 - \tilde{s}_q^\mu) \leq 2w_{\mathrm{med}}$, so the bonus rewards only media that stand out for the query, and its weight is fixed by w_{med}.

The mean does not have either property. Consider the weighted mean

$$s_{\mathrm{mean}}(e) = \frac{s_q^{\mathsf{txt}}(e) + w_{\mathrm{med}}\, \bar{s}^{\mathrm{med}}(e)}{1 + w_{\mathrm{med}}}, \qquad \bar{s}^{\mathrm{med}}(e) = \frac{1}{|\Sigma_e|} \sum_{\mu \in \Sigma_e} s_q^\mu(e). \tag{15}$$

It satisfies $s_{\mathrm{mean}}(e) < s_q^{\mathsf{txt}}(e)$ exactly when $\bar{s}^{\mathrm{med}}(e) < s_q^{\mathsf{txt}}(e)$, which the gap makes the typical case, and it falls by $w_{\mathrm{med}} \bar{\Delta}_e/(1 + w_{\mathrm{med}})$ when every gap grows by Δ_μ, with $\bar{\Delta}_e$ the average of Δ_μ over Σ_e. It therefore demotes media-bearing edges below text-only edges of equal text score, by an amount set by the encoder rather than by the evidence.

In summary, the median-anchored bonus only adds evidence and cancels any constant cross-modal offset. □

B.3 Proof of Proposition 3

Proposition 3. *Consolidation never deletes, and only repeated or revised facts change recall.*

Proof. Never deletes. The three cases of Eq. (10) are the only operations that modify the store after it is created, and nothing is archived or pruned. MERGE adds no edge and removes none, absorbing e^+ into e and raising a_e; SUPERSEDE and INSERT add e^+ to $\mathcal{E}^{(1)}$, and SUPERSEDE removes e from $\mathcal{L}$ but not from $\mathcal{E}^{(1)}$. By induction over writes, $\mathcal{E}^{(1)}$ after any sequence of writes contains every edge that was ever inserted. A superseded edge also stays reachable: by Eq. (11), $e \in \mathcal{L}_q$ whenever $\chi(c_q) = 1$.

Changes recall only on a repeat or a revision. Suppose every write of a benchmark falls into the INSERT case. Then $\mathcal{E}^{(1)}$, $\mathcal{E}^{(2)}$ and $\mathcal{L} = \mathcal{E}^{(1)}$ coincide with those of the append-only store $\mathcal{H}_{\mathrm{append}}$ in which consolidation is disabled, so $\mathcal{L}_q = \mathcal{E}^{(1)}$ for every q and the reweighting in Eq. (11) acts on no edge. The counters a_e enter none of Eqs. (5)–(8), which depend on the store only through $\mathcal{E}^{(1)}$, $\mathcal{E}^{(2)}$ and $\mathcal{L}_q$; hence

$$\textsc{Retrieve}(\mathcal{H}, q) = \textsc{Retrieve}(\mathcal{H}_{\mathrm{append}}, q) \qquad \text{for every question } q. \tag{16}$$

Contrapositively, recall can differ from the append-only store only if some write was a MERGE, which requires identical files and either identical text or agreeing values in a near-identical text (a repeat), or a SUPERSEDE, which requires a shared measure whose value changed (a revision).

Repeats do not grow the store. Write $\#_{\mathrm{ins}}$, $\#_{\mathrm{sup}}$ and $\#_{\mathrm{mrg}}$ for the numbers of writes in each case. An insert adds one edge to $\mathcal{E}^{(1)}$ and to $\mathcal{L}$, a supersede adds one to $\mathcal{E}^{(1)}$ and exchanges one in $\mathcal{L}$, and a merge adds none, so

$$|\mathcal{E}^{(1)}| = \#_{\mathrm{ins}} + \#_{\mathrm{sup}}, \qquad |\mathcal{L}| = \#_{\mathrm{ins}}. \tag{17}$$

A repeated fact therefore adds no edge, and never competes with its own copies for the k_1 slots.

Revisions rank below the current value. A superseded edge enters $\mathcal{L}_q$ only on a temporal cue, scaled by $\omega_{\mathrm{sup}} < 1$; as every kept edge clears the floor $\theta > 0$ of Eq. (7), it ranks strictly below a live edge of equal score. In summary, consolidation keeps every fact ever written, and alters recall only to compact repeats and to rank the current value first. □

C SUPERMEM ALGORITHM DETAILS

To illustrate the mechanism of SuperMem, we present its full workflow in Algorithm 1, comprising three phases that follow the three subsections of §3: **(1) Super-Hypergraph Construction with Conservative Consolidation.** Each unit becomes native nodes embedded by ϕ and bound by a first-level hyperedge, each stream a second-level one; every write is settled as a merge, supersede or insert by a deterministic rule that never deletes. **(2) Joint Cross-Modal Recall.** A question probes text and media nodes jointly, with a lexical term, a media bonus and a time discount, then narrows from the top k_2 streams to the top k_1 units. **(3) Sufficiency-Gated Delivery.** Whole units reach f only when the best score clears θ_{d}, as budgeted notes and raw files.

Algorithm 1 SuperMem: Omni-native Super-Hypergraph Memory

Require: Benchmark $\mathcal{D}$ of streams, question $q = (c_q, \mathcal{M}_q)$, encoder ϕ, frozen backbone f reading Σ_f
Ensure: Answer $\hat{y}_q$

// 1: Super-Hypergraph Construction with Conservative Consolidation (§3.1, §3.3)
Initialize $\mathcal{H} = (\mathcal{V}, \mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \leftarrow (\varnothing, \varnothing, \varnothing)$ and live edges $\mathcal{L} \leftarrow \varnothing$
for each stream $S \in \mathcal{D}$, with $E_S \leftarrow \varnothing$ **do**
  for each unit $x_t = (c_t, \mathcal{M}_t) \in S$ in order **do**
    Render and time: $\tilde{c}_t \leftarrow c_t \oplus r(\mathcal{M}_t)$; $\tau_{e_t} \leftarrow$ unit date, last date of S, or clip start
    Native nodes V_t by Eq. (2): $z_{v_t^\circ} = \phi(\tilde{c}_t)$, $z_{v_{t,j}} = \phi(\pi_{t,j})$, $z_{v_m} = \phi(m)$, one node per distinct file
    First-level hyperedge $e^+ \leftarrow V_t$ with summary vector $\bar{z}_{e^+}$ by Eq. (3)
    Candidates $\mathcal{N}(e^+)$ and slots $\Psi(\cdot)$ by Eq. (9); decide the case of e^+ by Eq. (10)
    if MERGE(e) **then**
      $a_e \leftarrow a_e + 1$ ▷ a repeat adds no edge
    else
      if SUPERSEDE(e) **then** $\mathcal{L} \leftarrow \mathcal{L} \setminus \{e\}$, e kept in $\mathcal{E}^{(1)}$ as superseded ▷ never deleted
      $\mathcal{E}^{(1)} \leftarrow \mathcal{E}^{(1)} \cup \{e^+\}$, $\mathcal{L} \leftarrow \mathcal{L} \cup \{e^+\}$, $E_S \leftarrow E_S \cup \{e^+\}$, $\mathcal{V} \leftarrow \mathcal{V} \cup V_t$
    end if
  end for
  Second-level hyperedge: $\mathcal{E}^{(2)} \leftarrow \mathcal{E}^{(2)} \cup \{E_S\}$ with $\bar{z}_{E_S}$ by Eq. (3)
end for
// 2: Joint Cross-Modal Recall (§3.2)
Probes $z_q \leftarrow \phi(c_q)$ and $P_q^\mu \leftarrow \{z_q\} \cup \{\phi(m) : m \in \mathcal{M}_q, \kappa(m) = \mu\}$; recall set $\mathcal{L}_q$ by Eq. (11)
for each $e \in \mathcal{L}_q$ **do**
  Joint score $s(e)$ by Eq. (5) ▷ text part $s_q^{\mathsf{txt}}(e)$ plus media bonus $\beta_q(e)$
  Time discount $\hat{s}(e) \leftarrow \rho_q(e)\, s(e)$ by Eq. (6); **if** $e \notin \mathcal{L}$ **then** $\hat{s}(e) \leftarrow \omega_{\sup}\, \hat{s}(e)$
end for
Stream scores $g(E_S)$, candidates $\mathcal{C}_q \leftarrow \mathcal{L}_q \cap \bigcup_{E_S \in \mathrm{Top}_{k_2}(g)} E_S$ by Eq. (7) ($\mathcal{C}_q \leftarrow \mathcal{L}_q$ if $|\mathcal{E}^{(2)}| \le k_2$)
Seeds $\mathcal{K}_q^0 \leftarrow \mathrm{Top}_{k_1}(\hat{s})$ over $\mathcal{C}_q$; one-hop bonus $s'(e)$; $\mathcal{K}_q \leftarrow \mathrm{Top}_{k_1}\{e \in \mathcal{C}_q : s'(e) \ge \theta\}$
if f is text-only (or lacks a stored modality) **then** keep the top k_{txt} (or k_{br}) edges of $\mathcal{K}_q$
// 3: Sufficiency-Gated Delivery (§3.2)
if $\mathcal{K}_q = \varnothing$ **or** $s_q^\star = \max_{e \in \mathcal{K}_q} s'(e) < \theta_{\mathrm{d}}$ **then**
  return $\hat{y}_q = f(c_q, \mathcal{M}_q, \varnothing)$ ▷ insufficient evidence: send nothing
end if
Notes $\mathcal{T}_q$: units $\tilde{c}_e$ in rank order within B characters, top n_{full} whole, the rest as ℓ_{w}-char windows
Raw files $\mathcal{F}_q \leftarrow$ FILL$\big(\bigcup_{e \in \mathcal{A}_q} \mathcal{M}_e\big)$ with $\mathcal{A}_q$ by Eq. (8)
return $\hat{y}_q = f(c_q, \mathcal{M}_q, \mathcal{T}_q \cup \mathcal{F}_q)$

Complexity Analysis. SuperMem involves three computational components corresponding to the phases. First, writing a unit calls no LLM beyond the shared renderings and makes only $1 + J_t + |\mathcal{M}_t|$ cached encoder calls; consolidation compares the new edge with the live edges of its own stream, costing $O(|E_S|d)$, so building $\mathcal{H}$ takes $O\big(\sum_t (1 + J_t + |\mathcal{M}_t|)\big)$ encoder calls and $O(|\mathcal{E}^{(1)}| \max_S |E_S|\, d)$ arithmetic, with a hash-shared file embedded once. Second, recall scores every node of the live edges with a single matrix product, $O(|\mathcal{V}|\, d)$ per question, plus BM25 over the stored texts and $O(|\mathcal{E}^{(2)}|\, d)$ for the stream level; the one-hop bonus touches only the k_1 seeds. Third, delivery sorts at most k_1 edges and sends at most B characters and a capped file payload, so the backbone context is bounded independently of the size of $\mathcal{H}$. Each component is deterministic and parallel across streams, and all vectors are cached, so a second pass costs no encoder call.

D BENCHMARK DETAIL

This appendix expands §4.1. All nine datasets are public releases used as is, with their own questions and gold answers; an *episode* is one ingest stream with its questions.

D.1 DATASET DETAILS

Table 5: **Evaluation suite**: 128 test questions each; streams are means over episodes; Types are the benchmark's own.

	Dataset	Ref.	Episode	#Ep.	Stream per episode	Types	Answer	Metric
A	LoCoMo	Maharana et al. (2024)	conversation	10	27 sessions, 73k chars	5	open	F1
	LongMemEval	Wu et al. (2025)	chat history	128	50 sessions, 490k chars	6	open	Acc
	MemoryAgentBench	Hu et al. (2026b)	long context	14	646k chars	4	open	Acc
B	Mem-Gallery	Bei et al. (2026)	persona dialogue	20	12 sessions, 50 images	9	open / image	F1
	MemEye	Guo et al. (2026)	scenario dialogue	8	28 sessions, 55 images	12	4-way choice	Acc
	LongDocURL	Deng et al. (2025)	document	39	93 page images	3	open	F1
C	M3-Bench-robot	Long et al. (2025)	robot video	17	31.5 min, video + audio	5	open	Acc
	OmniVideoBench	Li et al. (2026a)	video	108	20.7 min, video + audio	13	4-way choice	Acc
	LVOmniBench	Tao et al. (2026)	long video	19	33.4 min, video + audio	9	4-way choice	Acc

(A) Text. LoCoMo asks single-/multi-hop, temporal, open-domain and adversarial questions over long dialogues; LongMemEval gives each question its own fifty-session history; MemoryAgentBench tests retrieval, test-time learning, long-range understanding and conflict resolution. **(B) MM.** Mem-Gallery interleaves persona dialogues with photographs; MemEye puts the decisive evidence in the pixels, and our sample covers all twelve cells of its granularity (X1–X4) by reasoning-depth (Y1–Y3) grid; LongDocURL spans 52–142-page documents as 768 px renders. **(C) Omni.** M3-Bench-robot asks timestamped cross-modal questions over half-hour robot scenes; OmniVideoBench marks audio and visual evidence spans per answer; LVOmniBench videos run 20–53 minutes, with music and sound questions only audio answers.

D.2 EVALUATION DETAILS

Write-then-query. Per episode the store is emptied, the whole stream is written with retrieval off, the store is frozen, and questions are answered from retrieval alone. Units are a session with its images, a page, or a two-minute clip with its frames, audio and subtitles; the query prompt holds only the question, its own assets and the recall. All rows share the backbone at temperature 0, one planner→solver skeleton with 1,600 tokens per agent, and the embedder of §4.1; a reply with no answer is re-answered once.

Metrics. With $\mathrm{tok}(\cdot)$ splitting a lower-cased answer stripped of articles and punctuation, LoCoMo, Mem-Gallery and LongDocURL use SQuAD-style token F1 over gold aliases $\mathcal{Y}^*$,

$$\mathrm{F1}(\hat{y}, \mathcal{Y}^*) = \max_{y \in \mathcal{Y}^*} 2\,|\mathrm{tok}(\hat{y}) \cap \mathrm{tok}(y)| \,/\, \big(|\mathrm{tok}(\hat{y})| + |\mathrm{tok}(y)|\big) \in [0,1], \tag{18}$$

and the rest use accuracy: a choice reply is mapped onto the shown options (letter, text, or both score alike); an open reply hits on a normalised exact or numeric match or a contained gold alias of four or more characters. Abstention scores only where the benchmark defines it. A cell is the item mean $\times 100$, **ctx** the median recalled characters, and rows are compared by exact paired tests (McNemar; sign test for F1).

Table 6: **Reference controls** (`gemini-2.5-flash-lite`; subscript: gap to No memory; n/a: not annotated or over the 30 MB image payload). *Full context* replays the stream, and SuperMem beats it on five of seven columns; *Placebo*, wrong-episode recall, costs up to 21.8 points, hence SuperMem's gate; *Oracle* is the gold evidence.

Control	**A** Text			**B** MM			**C** Omni			
	LoCoMo F1	LMEval Acc	MemAB Acc	MemGal F1	MemEye Acc	LDocURL F1	M3B-r Acc	OmniVB Acc	LVOmni Acc	**ctx** chars
No memory same pipeline, query only	21.0	7.0	2.3	22.3	44.5	17.3	3.9	21.1	22.7	0
Full context whole stream in prompt	49.7$_{+28.7}$	35.9$_{+28.9}$	28.9$_{+26.6}$	n/a	n/a	55.4$_{+38.1}$	22.7$_{+18.8}$	21.1$_{\pm 0.0}$	24.2$_{+1.5}$	0
Placebo memory shuffled, token-matched	17.4$_{-3.6}$	10.2$_{+3.2}$	1.6$_{-0.7}$	18.8$_{-3.5}$	22.7$_{-21.8}$	4.7$_{-12.6}$	3.1$_{-0.8}$	14.1$_{-7.0}$	19.5$_{-3.2}$	2,519
Oracle memory gold evidence	51.5$_{+30.5}$	39.1$_{+32.1}$	n/a	56.2$_{+33.9}$	73.4$_{+28.9}$	69.3$_{+52.0}$	26.6$_{+22.7}$	37.5$_{+16.4}$	n/a	3,060

E BASELINE DETAIL

This appendix expands the rows of Table 2 into a per-method specification. For each of the 13 memory baselines, SuperMem and its three ablations we give three blocks: *(i) Overview*, the mechanism that defines the method and the store it builds; *(ii) Deployment*, how it is written and read under the protocol of Eq. 1; and *(iii) Hyperparameters*, the values used in every run. Rows follow Table 2, grouped by the two chips (§E.1–§E.7); Table 7 summarises them. Throughout, ϕ is the shared encoder, f the frozen backbone, $r(m)$ the shared rendering of a file (App. A.1), θ=0.35 the similarity floor every dense row applies, and *top-k* ranks stored entries by $\cos(\phi(c_q), \cdot)$ unless stated.

Table 7: **Design of every row of Table 2**: the stored form of a unit, LLM calls at write, recall rule and size, and whether raw files return to f.

Method	Stored per unit	LLM at write	Recall score	Recalled	Files
No memory	nothing	–	–	–	–
T BM25†	raw unit, term counts	none	Okapi BM25	top-5 windows	–
T MemGPT / Letta	main context + archival pages	eviction choice	main ∪ top-k archival	main + top-4	–
T Mem0	atomic facts	extract, ADD/UPD/DEL/NOOP	dense	top-5 facts	–
T Zep / Graphiti	fact edges with validity	edges, invalidation	dense over valid edges	top-5 edges	–
T A-Mem	linked notes	note, neighbour evolution	dense	top-5 notes	–
T Mem0^g	entity–relation triples	extract, ADD/UPD/DEL/NOOP	dense	top-5 triples	–
T G-Memory	task node + insights	insight distillation	dense seed + 1-hop	1 node + 3 insights	–
T MemOS	plaintext MemCube	none	dense + heat bonus	top-5 cubes	–
T HyperMem†	n-ary fact hyperedges	n-ary extraction	max over edge nodes	top-5 facts	–
M V-Mem	unit with modality lanes	none	lane-routed dense	top-5 units	✓
M MAGE	unit + text surrogates	none	dense on surrogate text	top-1 unit	–
O Omni-SimpleMem	summary + raw unit	summary	dense ∪ BM25, budgeted	top-5 units	✓
O M3-Agent	one binary edge per file	none	text + media node	top-1 edge	✓
O **SuperMem**	native nodes, 2-level hyperedges	none	joint, Eqs. 5–7	top-k_1, gated	✓

E.1 SHARED DEPLOYMENT

Harness. Every row implements the same two calls, WRITE and RETRIEVE, inside one harness, so rows differ only in what they store and how they rank. Write units are fixed across rows: a dialogue session with its images, a document page as its image, a 4,000-character block of a long context, or a 120 s video clip with three frames (320 px, one per 40 s), its mono 32 kbps audio and the subtitle cues inside it. Each episode's stream is written unit by unit before any question; recall is then placed in the prompt as "Recalled notes", followed by the question's own assets and the question, and answered by the planner→solver skeleton of App. D.2 with one re-answer for an empty reply.

Media path. A T row cannot hold a file, so each file is replaced at write time by its rendering $r(m)$, produced once by `gemini-2.5-flash-lite` ($\leq$ 160 tokens), cached per file and read identically by every row and backbone. M and O rows embed the files they keep with ϕ and may return raw files, capped for every row at 12 files and 12 MiB. Files a backbone cannot read are rendered by the same $r(\cdot)$ at query time (Table 8).

Table 8: **Backbones and encoder**, served through OpenRouter with the provider pinned; temperature 0 throughout. Price is $ per 1M input/output tokens.

Model	Native input	Context	Price	Answering budget	Rendered by $r(\cdot)$
`openai/gpt-oss-20b`	text	128K	0.03 / 0.13	reasoning *low*, 6,000 tokens	images, pages, frames, audio
`qwen/qwen3.7-flash`	text, image, video	1M	0.03 / 0.13	1,600 tokens; LVOmniBench as gpt-oss	audio
`google/gemini-2.5-flash-lite`	text, image, file, video, audio	1M	0.10 / 0.40	1,600 tokens	none
`google/gemini-embedding-2`	text, image, video, audio	–	–	d = 3072, one vector per item	–

E.2 TEXT-ADAPTED FLAT MEMORIES

No memory, the reference of every gap in Table 2, runs the same pipeline with an empty store: the prompt holds the question and its own assets only.

BM25†. *(i) Overview.* Okapi BM25 over the raw stream, the lexical floor: every unit is kept verbatim and ranked by term statistics, with no LLM and no encoder. *(ii) Deployment.* Units are tokenised to lower-cased alphanumerics; recall scores $\sum_{w \in c_q} \mathrm{idf}(w)\, \frac{\mathrm{tf}(w)(k_1+1)}{\mathrm{tf}(w)+k_1(1-b+b\,|x|/|\overline{x}|)}$ and returns, for each hit, the window of the unit densest in query terms. *(iii) Hyperparameters.* k_1=1.5, b=0.75, top-5, window 1,200 characters (stride 400).

MemGPT / Letta (Packer et al., 2024). *(i) Overview.* Virtual context management: a bounded main context that the agent curates itself and an unbounded archival store it pages from; under memory pressure the agent decides which notes stay in view. *(ii) Deployment.* Each unit enters main context; once main context exceeds its budget, f is shown the numbered notes and replies with those to keep, and every evicted note is embedded into archival storage. Recall returns the whole main context plus the top archival pages. *(iii) Hyperparameters.* Main context $3{,}000$ characters and 8 slots, archival top-4, eviction reply ≤ 60 tokens.

Mem0 (Chhikara et al., 2025). *(i) Overview.* Two phases: an LLM extracts the salient facts of an exchange, and each candidate fact is compared with its nearest stored facts so that an LLM decides ADD, UPDATE, DELETE or NOOP, keeping the store consolidated; a rolling summary conditions extraction. *(ii) Deployment.* Extraction names the speaker, resolves relative times against the session date and keeps numbers verbatim; the decision is read from a JSON reply; a deleted fact is tombstoned and hidden from recall. *(iii) Hyperparameters.* ≤ 12 facts per unit (unit truncated to $6{,}000$ characters), 5 nearest facts shown to the update step, summary ≤ 60 words, top-5.

E.3 Text-adapted Graph and Hypergraph Memories

Zep / Graphiti (Rasmussen et al., 2025). *(i) Overview.* A temporal knowledge graph whose edges carry a validity interval; a new fact that contradicts an old one invalidates it instead of overwriting it, so the store answers both what holds now and what held then. *(ii) Deployment.* Each unit is read into edges `subject | relation | object | date | fact`; the fact sentence is what is embedded and recalled, the triple and date what the bookkeeping uses. The nearest valid edges are shown to f, which names those the new edge invalidates; they stay in the store marked superseded. *(iii) Hyperparameters.* 4 nearest valid edges per check, top-5 valid edges at recall.

A-Mem (Xu et al., 2025). *(i) Overview.* Zettelkasten notes: each note receives keywords and is linked to its nearest notes, and the linked notes are themselves rewritten in light of the new one (memory evolution). *(ii) Deployment.* A unit is written as atomic notes, one fact per line with its keywords; its 3 nearest notes are linked and passed to f with the new note for rewriting. *(iii) Hyperparameters.* Link k=3, note ≤ 220 tokens, evolution ≤ 320 tokens, top-5.

Mem0^g (Chhikara et al., 2025). *(i) Overview.* The graph variant of Mem0: the stored unit is an entity–relation triple rather than a free-text fact, and the same update loop consolidates the graph. *(ii) Deployment.* Identical to Mem0 except that extraction emits `subject | relation | object` lines and only parsed triples are kept, so the pair of rows isolates the effect of the graph schema. *(iii) Hyperparameters.* As Mem0.

G-Memory (Zhang et al., 2025). *(i) Overview.* A hierarchical graph memory: a task layer links similar interactions, and an insight layer distils lessons over each neighbourhood; recall expands from the best node through its neighbours and attaches supported insights. *(ii) Deployment.* Each unit is a node linked to every node within the edge threshold; a node landing in a neighbourhood of at least two nodes triggers one distilled insight attached to its members. Recall takes the best node, expands one hop and returns the insights whose support meets the reached set. *(iii) Hyperparameters.* Edge threshold 0.7, seed top-1, hop 1, insights top-3 (≤ 90 tokens each).

MemOS (Li et al., 2025). *(i) Overview.* A memory operating system whose MemCubes carry provenance and a heat counter; hot cubes are promoted and cold ones decay, and recall reads the promoted set first. We run its plaintext MemCube tier. *(ii) Deployment.* Each unit is one cube; a cube read enough times becomes hot and outranks a cold cube of equal similarity; a periodic sweep cools every cube. *(iii) Hyperparameters.* Promotion after 2 reads, hot bonus 0.05, decay sweep every 400 writes, top-5.

HyperMem† (Yue et al., 2026a). *(i) Overview.* A text hypergraph memory for long conversations: an exchange is read into n-ary facts, each stored as one hyperedge over all the entities it binds, so a fact is reachable from any of its members. *(ii) Deployment.* Extraction writes each fact as `entity; ...; entity :: fact`; each entity and the fact sentence receive a vector, and an edge scores by its best-matching vector. *(iii) Hyperparameters.* Max arity 6, extraction ≤ 500 tokens, top-5.

E.4 VISION-NATIVE MEMORIES

V-Mem (Kang et al., 2026). *(i) Overview.* Modality-routed retrieval: to cross the modality and similarity–relevance gaps, a query is routed to the lane of the modality that holds its evidence and matched within that lane, returning the whole round. *(ii) Deployment.* Each unit keeps its text vector and native image and audio vectors from ϕ; the target lane is read from the query; a text query seeking an image or a sound searches with an LLM-written hypothetical caption, and a query carrying an image adds keywords read from it. Recalled units return their raw files. *(iii) Hyperparameters.* ≤ 8 media vectors per modality and unit, anchor ≤ 60 tokens, top-5, text-lane fallback when a lane is empty.

MAGE (Feng et al., 2026). *(i) Overview.* A hypergraph memory that keeps modality pointers alongside extracted textual surrogates rather than native vectors. *(ii) Deployment.* Each unit is one hyperedge tagged with its modalities; every file is replaced by an LLM description with the modality prompts of App. A.1, and retrieval embeds the unit text with its surrogates. *(iii) Hyperparameters.* ≤ 8 files per modality, surrogate ≤ 160 tokens, top-1.

E.5 NATIVE ALL-MODAL MEMORIES

Omni-SimpleMem (Liu et al., 2026a). *(i) Overview.* An all-modal memory built on selective ingestion (a unit is stored only if novel), atomic units whose searchable summary is separated from the raw content, and progressive retrieval under a token budget over dense and BM25 candidates. *(ii) Deployment.* A unit whose vector is too close to a stored one is skipped; a stored unit is prefixed by a one-sentence summary and keeps its native media vectors; recall unites the dense list with the BM25 list and admits units while the budget lasts, returning their raw files. *(iii) Hyperparameters.* Novelty threshold 0.93, dense top-10 $\cup$ BM25 top-5, budget $6{,}000$ characters, top-5.

M3-Agent (Long et al., 2025). *(i) Overview.* A multimodal agent memory that turns long video into an entity-centric graph whose edges link an entity or a clip to one observation at a time. *(ii) Deployment.* Each unit is written as one binary edge per file, pairing the unit text with that file's native vector from ϕ, on the two-minute clips that are also M3-Agent's own unit; recall scores an edge as the mean of its text and media similarity and returns its file. *(iii) Hyperparameters.* ≤ 8 files per modality and unit, weights $w_{\text{text}}{=}w_{\text{media}}{=}1$, top-1.

E.6 SUPERMEM

(i) Overview. §3 and Algorithm 1. *(ii) Deployment.* Writing embeds the nodes of Eq. 2 and calls no LLM; recall follows Eqs. 5–8 and consolidation Eqs. 9–11. One frozen configuration, hashed in every run summary, serves every column and backbone; on `qwen3.7-flash` × LVOmniBench, run by every row with the capped reasoning of Table 8, all k_1 kept edges are delivered (k_{br} off). *(iii) Hyperparameters.* Table 9.

Table 9: **SuperMem hyperparameters**, symbols of §3.

Knob	Value	Knob	Value
k_1 edges / k_2 streams kept	8 / 8	B note budget / n_{full} / ℓ_{w} (chars)	70,000 / 5 / 1,200
η stream mix / δ one-hop bonus	0.5 / 0.05	ℓ passage / J_t cap / unit key (chars)	1,500 / 12 / 6,000
θ floor / θ_{d} gate	0.35 / 0.45	media nodes per modality and edge	8
w_{lex} / b_0 / min. BM25 text (chars)	0.2 / 5.0 / 200	n_{e} / γ / θ_{file}	3 / 0.25 / 0.45
media term / w_{med}	bonus / 1.0	FILL files / bytes	12 / 12 MiB
$T_{1/2}$ / ω_{fut} / ω_{sup}	8.0 / 0.5 / 0.9	k_{txt} / k_{br}	5 / 1
θ_{dup} / $\lvert\mathcal{N}(e^+)\rvert$	0.97 / 3	arity / $r(m)$ in $\tilde{c}_t$	unbounded / on

E.7 ABLATIONS

Each row of Table 3 is the frozen configuration with one switch, on `gemini-2.5-flash-lite` with the same units, encoder and prompts; accuracy averages three draws. **w/o omni-native nodes** (`surrogate` = on): no media node is written, a file enters only as $r(m)$ in $\tilde{c}_t$, and no raw file is returned, as in MAGE. **w/o joint cross-modal recall** (`joint` = off): media nodes and files are kept, but Eq. 5 loses its media term. **w/o conservative consolidation** (`consolidate` = off): every write is an INSERT. The last two columns write each stream of LoCoMo, Mem-Gallery and MemEye as its units repeated eight times, against a single write.

F EXPERIMENT DETAIL

This appendix gives the protocol and per-benchmark results behind §4.2–§4.9. Unless stated otherwise, all analyses use `gemini-2.5-flash-lite`, the frozen configuration (Table 9) and the 128 test questions per benchmark (Table 5). *Accuracy* comes from full runs (those of Table 2, plus fresh runs for RQ2 and RQ6), *retrieval metrics* from their recall records, and *offline replays* (RQ5, RQ7) rebuild a store and rerun recall with all vectors cached and every LLM call disabled; as writing and recall call no LLM, a replay reproduces a run's recall exactly, without sampling noise. **R@5** is the mean fraction of a question's gold evidence in its top five units, and **Own@1** the share of questions whose top unit comes from the episode asked about; retrieval means are over the five media sets with gold evidence, plus LVOmniBench for Own@1.

F.1 MAIN RESULTS (RQ1)

Setup. Figure 4 averages each benchmark of Table 2 over the three backbones, scales it by the best row, and draws the strongest row of each store family: MemOS, V-Mem and M3-Agent. **Results.** Per block (Table 10), SuperMem leads all three, most widely on MM, while the runner-up changes from block to block: no baseline family is strong on every kind of evidence, and the omni-native super-hypergraph is.

Table 10: **Block averages** over three backbones; underline: best baseline.

Row	Text	MM	Omni	Avg.
No memory	9.2	21.9	16.9	16.0
T MemOS	26.7	33.1	18.7	26.2
M V-Mem	32.8	30.5	20.6	28.0
O M3-Agent	22.9	30.5	21.5	25.0
O **SuperMem**	**35.6**	**50.1**	**26.7**	**37.5**

F.2 ABLATION STUDY (RQ2)

Setup. The variants (App. E.7) and the full model were each run three times on the six MM and Omni benchmarks; Table 3 reports the mean block accuracy. On the Text block the first two switches have nothing to act on, as its streams hold no files, and the third changes no recall anywhere: with consolidation off, all 1,152 questions keep exactly the full model's units. **Results.** Per benchmark (Table 11), both switches lower R@5 on every set not saturated at 100 and Own@1 on all six sets, and neither variant is above the full model anywhere; the largest drops are 8.0 R@5 (OmniVideoBench) and 10.2 Own@1 (Mem-Gallery) without omni-native nodes, and 9.5 R@5 and 12.5 Own@1 (Mem-Gallery) without joint recall.

Table 11: **Per-benchmark retrieval** (RQ2, RQ4). Rescued / lost: gold units joint recall adds to / removes from the top five; cosines: medians over the top eight units; n/a: not applicable.

		MemGal	MemEye	LDocURL	M3B-r	OmniVB	LVOmni	**Mean**
R@5	SuperMem	**93.3**	**83.9**	**74.9**	**100.0**	**80.7**	n/a	**86.6**
	w/o omni-native nodes	88.2	80.2	69.4	100.0	72.7	n/a	82.1
	w/o joint recall	83.8	80.2	69.4	100.0	73.7	n/a	81.4
Own@1	SuperMem	**84.4**	**100.0**	**96.1**	**69.5**	**64.8**	**55.5**	**78.4**
	w/o omni-native nodes	74.2	95.3	90.6	64.1	56.3	50.0	71.7
	w/o joint recall	71.9	95.3	90.6	64.1	55.5	50.0	71.2
Gold units	rescued by joint recall	41	19	15	0	27	n/a	102 (sum)
	lost by joint recall	1	5	1	0	2	n/a	9 (sum)
Median cos.	text node	0.61	0.65	0.63	0.60	0.60	0.57	0.61
	image node	0.33	0.38	0.45	0.33	0.36	0.35	0.36
	audio node	n/a	n/a	n/a	0.61	0.63	0.55	0.60

F.3 ANALYSIS OF THE OMNI-NATIVE SUPER-HYPERGRAPH (RQ3)

Setup. A unit's arity is the number of nodes its first-level hyperedge binds: its text passages and media nodes (at most eight per modality). The sweep reruns the pipeline with a cap $k \in \{2, 3, 4\}$ that splits every larger hyperedge into pieces of at most k nodes; k=2 is the pairwise restriction of Proposition 1, the surrogate restriction is the w/o-omni-native-nodes variant, and MAGE and M3-

Agent are the rows of Table 2. **Results.** Native units are high-order (Table 12): a Mem-Gallery session binds 9.6 nodes on average and up to 21, a video clip 5.5–6.5. R@5 never falls as k grows, and the gain tracks the arity: largest on OmniVideoBench and MemEye, and nil only where a unit binds two nodes (a LongDocURL page) or recall is already perfect. SuperMem matches or beats every special case on every set and leads MAGE and M3-Agent by 33.4 and 41.2 mean R@5.

Table 12: **Stores and designs** (RQ3). Nodes per unit: mean (max); R@5 of SuperMem under the arity cap k (k=2: pairwise), of its surrogate restriction and of MAGE and M3-Agent.

				SuperMem, arity cap k				**Special cases**		
Dataset	**Units**	**Ep.**	**Nodes/unit**	k=2	3	4	∞	surrogate	MAGE	M3-Agent
Mem-Gallery	240	20	9.6 (21)	86.0	89.3	91.5	**93.3**	88.2	68.2	47.3
MemEye	221	8	3.8 (13)	67.4	72.5	74.5	**83.9**	80.2	42.4	11.6
LongDocURL	3,631	39	2.0 (2)	**74.9**	**74.9**	**74.9**	**74.9**	69.4	40.4	47.3
M3-Bench-robot	276	17	6.5 (8)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	85.9	89.1
OmniVideoBench	690	108	5.8 (7)	61.5	68.6	71.1	**80.7**	72.7	29.1	31.7
Mean				78.0	81.1	82.4	**86.6**	82.1	53.2	45.4

F.4 Joint Cross-Modal Recall across the Modality Gap (RQ4)

Setup. For every question of the six media sets we read the per-modality similarity of the question to the nodes of its top eight units, as logged by Eq. 5. A gold unit is rescued if it is in the top five of the full model only, and lost if only in that of the w/o-joint-recall variant. **Results.** The gap is a property of the modality, not of the benchmark (Table 11): image nodes sit 0.18–0.28 below text nodes on every set, while audio stays within 0.03, hence each modality is compared with its own median. So calibrated, joint recall raises Own@1 on all six sets and rescues 11 gold units per one lost.

F.5 Conservative Consolidation under Repetition (RQ5)

Setup. A replay writes each stream of LoCoMo, Mem-Gallery and MemEye $f \in \{1, 2, 4, 8\}$ times in sequence, with consolidation on and off (append-only). **Results.** SuperMem merges every repeat into the edge it repeats, so its store and R@5 equal those of a single write (Table 13). The append-only store grows linearly, and as each copy scores like its original, copies of the best unit displace other gold units: R@5 falls monotonically, by 26.4 on average at f=8. With consolidation on, the other two ablations are equally unaffected, so the guarantee does not depend on how units are scored.

Table 13: **Repeated writes** (RQ5): edges and R@5 with every stream written f times; SuperMem's are constant in f.

	SuperMem, any f		**Append-only edges**				**Append-only R@5**			
Dataset	Edges	R@5	f=1	2	4	8	f=1	2	4	8
LoCoMo	272	**88.2**	272	544	1,088	2,176	88.2	85.7	77.8	61.5
Mem-Gallery	240	**93.3**	240	480	960	1,920	93.3	89.3	84.5	74.2
MemEye	221	**83.9**	221	442	884	1,768	83.9	71.2	65.3	50.3

F.6 Efficiency: Accuracy at Matched Context (RQ6)

Setup. The sweep reruns SuperMem on all nine benchmarks with note budget $B \in \{2\text{k}, 4\text{k}, 8\text{k}\}$ characters (frozen: 70k); context is the median recalled characters over all 1,152 questions and the score the Avg. of Table 2. Evidence density cuts each recalled text of the runs of Table 2 to its first b characters and counts the gold evidence it still holds, over the seven benchmarks with gold evidence and the eight baselines whose records keep the unit headers the matcher reads. **Results.** SuperMem averages 31.9, 35.2 and 36.7 at 2,114, 3,699 and 6,587 recalled characters, and 39.9 at the frozen budget (17,307), against 30.1 for V-Mem on 5,480: it does not need context to lead, because it ranks the evidence first. Its first 500 characters beat any baseline's first 32,000 in gold evidence (Table 14).

Table 14: **Evidence density** (RQ6): gold evidence (%) within the first b recalled characters.

Row	b=500	1k	2k	4k	8k	16k	32k
BM25	23.9	26.2	31.2	33.9	34.4	34.4	34.4
MemGPT / Letta	15.8	16.1	16.1	34.3	38.6	41.5	41.5
G-Memory	45.5	45.5	45.5	45.5	45.5	45.5	45.5
MemOS	40.4	42.3	44.3	46.5	51.5	55.4	57.8
V-Mem	40.4	41.3	42.6	47.0	51.1	54.1	55.6
MAGE	48.0	48.0	48.0	48.0	48.0	48.0	48.0
Omni-SimpleMem	31.7	32.8	33.3	36.3	38.1	38.1	38.1
M3-Agent	42.5	42.5	42.5	42.5	42.5	42.5	42.5
SuperMem	**58.5**	**61.2**	**63.3**	**68.1**	**71.2**	**75.3**	**78.2**

F.7 Robustness and Sensitivity (RQ7)

Setup. The sweep replays recall on the five media sets, moving one knob at a time from the frozen configuration (edges kept k_1, streams kept k_2, weights w_{lex} and w_{med}), with joint and with text-only scoring. The paired tests compare SuperMem with each baseline question by question under each benchmark's metric: the win rate is wins over wins plus losses, p the two-sided sign test on non-tied questions, and each of the 58 task categories with at least five questions counts as won or tied when SuperMem's mean is at least the baseline's. **Results.** The frozen k_1 and k_2 lie on plateaus within 0.3 R@5 of their neighbours, w_{lex} at its peak and w_{med} on a rising curve, so no result hinges on a finely tuned value, and joint recall beats text-only scoring at all 21 settings, by 2.3–7.9 (Table 15). SuperMem wins more questions than it loses against every baseline (Table 16), from 64.2% against V-Mem to 83.5% against Zep, and every 95% lower bound is above 59%.

Table 15: **Sensitivity** (RQ7): mean R@5 over five media sets, joint / text-only; shaded: frozen.

Knob	**Value**: joint / text-only					
k_1	2: 74.7 / 66.8	4: 84.0 / 78.2	6: 86.9 / 81.4	8: 86.6 / 81.4	12: 86.6 / 81.4	16: 86.7 / 81.4
k_2	2: 86.8 / 83.4	4: 86.3 / 81.6	8: 86.6 / 81.4	16: 86.6 / 81.2	32: 86.6 / 81.2	
w_{lex}	0.0: 84.8 / 77.0	0.1: 85.4 / 80.2	0.2: 86.6 / 81.4	0.3: 85.0 / 81.5	0.4: 84.5 / 80.7	
w_{med}	0.25: 83.7 / 81.4	0.5: 85.1 / 81.4	1.0: 86.6 / 81.4	1.5: 87.1 / 81.4	2.0: 88.1 / 81.4	

Table 16: **Paired tests** (RQ7): W/L/T, win rate (%), categories won or tied (of 58); all $p < 10^{-8}$.

Baseline	W	L	T	Win	Cat.	Baseline	W	L	T	Win	Cat.
T BM25	316	119	717	72.6	51	T MemOS	298	128	726	70.0	49
T MemGPT / Letta	306	116	730	72.5	46	T HyperMem	359	124	669	74.3	44
T Mem0	357	116	679	75.5	48	M V-Mem	276	154	722	64.2	43
T Zep / Graphiti	430	85	637	83.5	49	M MAGE	307	130	715	70.3	44
T A-Mem	399	101	652	79.8	51	O Omni-SimpleMem	303	124	725	71.0	50
T Mem0^g	392	98	662	80.0	53	O M3-Agent	293	144	715	67.0	46
T G-Memory	326	127	699	72.0	49						

F.8 Case Study (RQ8)

Setup. Every row of Table 4 is read from its record for M3-Bench-robot question `test_00025` in the runs of Table 2. `bedroom_07` is written as 16 first-level hyperedges, one per two-minute clip with its subtitle note, three frames 40 s apart and its audio, nested in the recording's second-level hyperedge; every row passes the files of its recalled units, in rank order and at most 12. **Results.** The subtitle reads the request as "get me some clue", so lexical matching on "glue" misses it; SuperMem ranks the clip first by its joint score (text 0.73, audio 0.63), 0.10 ahead of the next unit, and delivers its top n_e=3 units with their 12 files (Figure 11), whose frames show the backbone the seating. Every other row reports that the evidence lacks the answer. All eight units SuperMem keeps come from the recording asked about, and the clip with the highest audio similarity (06:00, 0.68) ranks last, as its text matches least: no single modality decides alone.

G Extended Related Work

G.1 Structured Agent Memory

Persistent memory couples storage with selective access: MemGPT (Packer et al., 2024), Mem0 (Chhikara et al., 2025), MemOS (Li et al., 2025), and PlugMem (Yang et al., 2026c) establish management interfaces, while LightMem (Fang et al., 2026a), SimpleMem (Liu et al., 2026b), Hindsight (Latimer et al., 2026), and EverMemOS (Hu et al., 2026a) emphasize compression, reflection, and consolidation. Beyond these interfaces, the graph-memory survey (Yang et al., 2026a) frames relational organization: GraphRAG (Edge et al., 2025) and A-Mem (Xu et al., 2025) support graph retrieval and associative notes, while G-Memory (Zhang et al., 2025), GAM (Wu et al., 2026), HiGram (Yue et al., 2026b), and xMemory (Hu et al., 2026c) organize access across abstraction levels. HyperGraphRAG (Luo et al., 2025), HGMEM (Zhou et al., 2026a), HyperMem (Yue et al., 2026a), and HyperSkill (Xu et al., 2026) extend higher-order structure to knowledge retrieval, working memory, conversations, and skills, respectively; EdgeMem (Cui et al., 2026) preserves source evidence in multi-anchor hypergraphs. Maintaining these structures requires both adaptation and compression: FluxMem (Fang et al., 2026b) and CABLE (Tan et al., 2026) evolve or extend connections; MRAgent (Ji et al., 2026), REALM (Song et al., 2026), and AutoViewMem (Cao et al., 2026b) reconstruct or reorganize memories; MemForest-Temporal (Chen et al., 2026b) and MemForest-EventTree (Wang et al., 2026a) use temporal indexing and event merging. Zep (Rasmussen et al., 2025), TOKI (Wang, 2026), SodaMem (Wan et al., 2026), MemoryLACE (Yacoubi et al., 2026), and LatticeMind (Zhou et al., 2026b) address temporal validity or conflicts; Selective Forgetting (Rusu et al., 2026) cautions that graph extraction can lose useful evidence. Complementing these update mechanisms, Router-Mem (Lin et al., 2026), MESA (Zhao et al., 2026), MemArbiter (Dong et al., 2026), and Jev-Mem (Jiang et al., 2026) condition memory access on evidence or budgets, while MemCalib (Cao et al., 2026a) evaluates inappropriate memory use. Where MACE (Yang et al., 2026b) organizes adaptive graphs of memory subgraphs, SuperMem contributes nested episode and evidence-unit hyperedges that bind native evidence, coupling hierarchical recall with conflict-aware consolidation that preserves superseded evidence and its validity intervals.

G.2 Multimodal Agent Memory

Multimodal memory extends persistence from linguistic abstractions to perceptual experience: MIRIX (Wang & Chen, 2025), M3-Agent (Long et al., 2025), MemVerse (Liu et al., 2026c), TeleMem (Chen et al., 2026a), and Omni-SimpleMem (Liu et al., 2026a) develop long-term multimodal stores, with VoiceMem (Xie et al., 2026) and LightMem-Ego (Chen et al., 2026c) targeting speech and everyday experience. Retaining detail under growing histories motivates semantic refinement in ViLoMem (Bo et al., 2026) and video abstraction in WorldMM (Yeo et al., 2026), MM-Mem (Lian et al., 2026), and MemoryCard (Yang et al., 2026d). Cross-observation dependencies motivate complementary representations: VimRAG (Wang et al., 2026b) builds multimodal reasoning graphs, MAGE (Feng et al., 2026) uses temporal hypergraphs, GraphMemix (Li et al., 2026b) selects evidence forests, and EM^{2}Mem (Chen et al., 2026d) binds heterogeneous evidence to events; Collaborative VLM Memory (Zhang et al., 2026) discusses shared visual context and revision. These representations must also support usable readout: PMMC (Sun et al., 2026) precompiles memory, AdaMM (Tian et al., 2026) enables analytic operations, V-Mem (Kang et al., 2026) routes retrieval by modality, and MEMO (Gao et al., 2026) budgets textual and visual presentation. MMA (Lu et al., 2026) incorporates reliability and abstention, while ChronoProfiler (Zhou et al., 2026c) tracks temporally changing preferences on ReaLMem, emphasizing that retained evidence must remain relevant at decision time. SuperMem advances this connection between representation and readout by binding native text, image, audio, and video-frame embeddings within nested hyperedges, combining joint cross-modal scoring and whole-unit recall with deterministic sufficiency gating and budgeted delivery of original media.