INTRINSIC REPRESENTATION MEMORY: DECOUPLING STRUCTURAL TOPOLOGY AND SEMANTICS FOR LIFE-LONG AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equipping agents with long-term memory raises the challenge of representation entanglement. Storing encoder-specific vectors ties retrieval to a particular indexing model. We propose Intrinsic Representation Memory (IRM), a framework that decouples structural memory from semantic representation. Instead of generating text descriptions or storing frozen embeddings, IRM constructs a vector-free symbolic graph from a model's intrinsic attention salience during a standard forward pass. This zero-generation write retains entity links as inspectable symbolic memory. At read time, the current model reuses this graph and dynamically regenerates value representations within its own semantic space. IRM extends to multimodal settings, processing both text and raw image pixels without a separate captioner or embedder. Across four models, cross-built graphs retain at least 94.8% of each reader's own-graph text recall. With a common frozen 9B answer reader, IRM improves NarrativeQA F1 by 5.56 points over HippoRAG-2, with lower F1 on MuSiQue and mixed results on LoCoMo. In a two-memory study using the same Qwen model, construction is more than an order of magnitude faster than HippoRAG-2 under the measured implementations, while per-query latency is higher. By eliminating symbolic graph reconstruction during reader changes and accelerating memory writes, IRM offers a reader-portable memory foundation for continually evolving multimodal agents.

1 INTRODUCTION

Retrieval-augmented systems serve as long-term memory for large language model (LLM) agents. Ideally, such a memory is written once and stays useful as its reader is upgraded across generations. Many memory systems generate facts, relations, or summaries when processing new text (Chhikara et al., 2025; Gutiérrez et al., 2025; Edge et al., 2024; Sarthi et al., 2024). When these representations are embedded for retrieval, the resulting index is tied to the encoder, creating representation entanglement. Replacing the answer model leaves that index unchanged, while its stored representations remain tied to the indexing encoder. We ask whether a memory can instead decouple structural connectivity from semantic embedding, deferring continuous semantic representation to the evolving model that reads it.

Intrinsic Representation Memory (IRM)[1] uses a frozen model's internal representations in separate write and read stages (Fig. 1). IRM uses intrinsic attention salience to extract symbolic links at write time, capturing attention-derived associations without generating auxiliary text. Value-space matching and query likelihood score evidence at read time. The store preserves transferable topological connectivity while the current reader computes contextual relevance in its own latent space. This separation aims to combine cross-model portability, multi-hop retrieval, and competitive accuracy with a single frozen reader for continuously updating agents. We make three contributions.

[1] Anonymous code at https://anonymous.4open.science/r/IRM-73C0/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

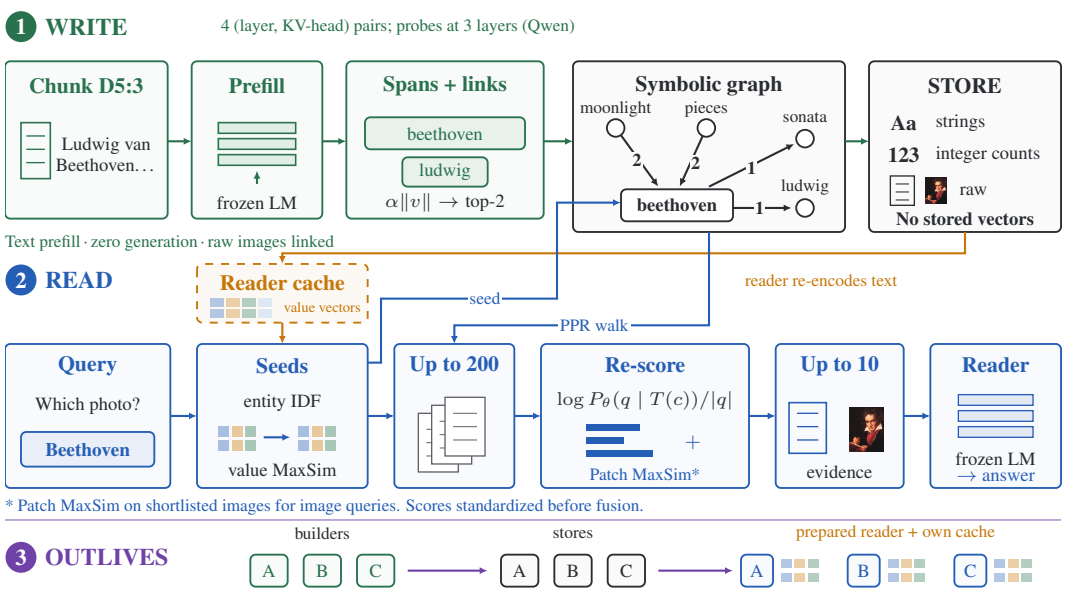


Figure 1: **Overview of IRM.** Writing produces a symbolic graph without generating extraction text. Reading uses the current model's representations, allowing the graph to outlive its builder. Graph arrows denote stored directed events, while personalized PageRank (PPR) traverses undirected links. LM denotes language model, KV key–value and IDF inverse document frequency. MaxSim aggregates the best token or patch matches.

C1. A zero-generation write driven by intrinsic salience. An agent runs a forward pass to read each chunk, and IRM constructs memory directly from that pass. A probe marks entity spans, and IRM uses the model's internal salience readout $\alpha \|v\|$ (attention weight times value norm) to retain each span's strongest neighbors as directed, counted edges. The write therefore generates zero tokens, calls no separate extraction LLM or embedder, and retains the resulting connectivity as inspectable symbolic memory.

C2. A reader-portable store that survives its builder. The vector-free graph records contain only surface spans and directed integer-count edges. Document frequencies are derived at load. The graph's topology reflects the builder's selected links, but it has no latent coordinate system tied to that specific architecture. A new reader regenerates its own value representations while reusing the symbolic store. Exact per-chunk counts also support reversible deletion of symbolic contributions.

C3. One reader for text and pixels. The deferred read serves both modalities through the current reader's own representations. The store retains raw evidence and symbolic structure. Un-pooled matching scores text candidates at corpus scale and compares shortlisted images against the query image using modality-specific comparison heads. The top-ranked evidence images then enter the reader as raw pixels, preserving visual details for interpretation by the current reader.

2 Related Work

Agent memories and lifelong learning. Many memory systems store encoder-specific vectors or generated textual abstractions at write time. The former depend on the indexing encoder. The latter preserve the writer's extraction or summarization choices. Dense retrieval-augmented generation (RAG) couples its index to a trained encoder (Lewis et al., 2020; Karpukhin et al., 2020), so changing the passage encoder generally requires re-encoding the indexed corpus. Mem0, GraphRAG, RAPTOR, A-Mem, and Zep generate facts, contextual descriptions, or summaries as they construct or update memory (Chhikara et al., 2025; Edge et al., 2024; Sarthi et al., 2024; Xu et al., 2025; Rasmussen et al., 2025). HippoRAG-2 traverses a graph built by open information extraction (OpenIE) with personalized PageRank (Gutiérrez et al., 2025). IRM retains seeded graph traversal while keeping generated extraction text and embeddings out of its symbolic store. This establishes a reusable

topological foundation for continuously updating agents, with semantic representations prepared by the current reader.

Generation-free retrieval and inspectable structure. Indexes without generation keep surface symbols alone or add a separate embedder for meaning. BM25 and named entity recognition (NER)-based co-occurrence graphs provide portable lexical or symbolic indexes (Robertson & Zaragoza, 2009; Wang, 2025). LinearRAG links entities, sentences, and passages without LLM relation extraction but retains a separate sentence embedder (Zhuang et al., 2026). Retrieval-head analysis identifies heads that recover information from long contexts (Wu et al., 2025). Frozen language models can re-rank passages through query likelihood (Sachan et al., 2022) or query-conditioned attention (Zhang et al., 2025; Chen et al., 2025). IRM uses inspectable internal salience readouts to construct symbolic links at write time, without generating extraction text. At read time, un-pooled value comparisons seed traversal of this portable graph. Graph traversal supplies candidates beyond direct seed matches, and unsupervised passage re-ranking (UPR) scores the resulting shortlist.

Multimodal retrieval. The following multimodal retrievers index model-specific image representations. ColPali applies MaxSim late interaction to stored patch vectors from a fine-tuned vision-language model (Faysse et al., 2025). FreeRet uses a frozen multimodal model for embedding search and subsequent re-ranking (Zhu et al., 2026). MuRAG learns multimodal retrieval and generation jointly (Chen et al., 2022), whereas VisRAG uses a trained retriever to index document pages as images (Yu et al., 2025). IRM performs token- and patch-level matching with values dynamically regenerated by its frozen reader, keeping latent vectors out of the symbolic store while preserving raw image evidence for interpretation by later readers.

3 METHOD

To decouple topological memory structure from continuous semantic representation, IRM uses separate write and read stages, each running its own model prefills with distinct scoring readouts. The write stage distills the reading pass into a reader-portable symbolic graph. The store keeps that graph with raw evidence and no model-specific vectors. The read stage lets the loaded reader recompute contextual relevance within its own representation space.

3.1 WRITE, THE GRAPH AS A BYPRODUCT OF READING

The write stage constructs symbolic nodes and edges from a single prefill pass, without generating extraction text or calling a separate extraction LLM. Before deployment, four (layer, KV head) pairs in softmax-attention layers are selected once per model by the area under the receiver operating characteristic curve (AUC) for retrieval on disjoint external HotpotQA training queries. We call these pairs *comparison heads*. Within each model, its own selected coordinates provide attention weights for writing and values for reading (Supp. A.2).

An entity probe identifies the nodes from the input hidden states at the distinct comparison-head layers. One logistic-regression direction per layer is fitted once per model on rule-generated proper-name labels from the external training passages, never on benchmark labels. A token's score is the mean of the probe logits across these layers, and detected spans are extended to whole words. Inference uses the fitted logits without a capitalization rule. Each node is stored as a normalized surface span, not a vector, with one token run retained per surface key in each chunk. No semantic alias resolution is applied.

The edges come from the attention heads that are already comparing tokens as the model reads. A head relates token i to token j by $\alpha_{ij} = \mathrm{softmax}_j(q_i^\top k_j/\sqrt{d} + M_{ij}^{\mathrm{attn}})$ with causal mask M^{attn} and moves information by $\sum_j \alpha_{ij} v_j$. *Within* one chunk's context the full operator applies, and its native salience readout is

$$a_{i\to j} \;=\; \alpha_{ij}\,\|v_j\|, \tag{1}$$

the norm of token j's weighted value contribution before the head output projection. Acting as an inspectable proxy for attention-mediated associations, this readout measures weighted value contributions between tokens. Aggregating these contributions between entity spans provides one implementation of generation-free symbolic link selection, without a separate relation extractor. This is related to norm-based attention analysis (Kobayashi et al., 2020), with the probe and the use of

this readout as the structural backbone for graph edges introduced here. Eq. 1 is computed for one attention head. Under grouped-query attention, v_j is the value of the KV head through which query head h attends. Averaging Eq. 1 over all query heads in the four selected KV groups gives $\bar{a}_{i\to j}$. For entity spans u and v, the mean salience and event count are

$$w(u{\to}v) = \mathrm{mean}_{u\times v}\, \bar{a}_{i\to j}, \qquad c(u{\to}v) = \sum_{i\in u, j\in v} b_i\, \mathbf{1}[j \neq 0,\ \bar{a}_{i\to j} > \mu_i{+}\sigma_i], \tag{2}$$

where the mean and count range over $(i, j) \in u \times v$, after zeroing column 0. μ_i, σ_i are token i's salience mean and standard deviation over preceding positions excluding position 0. The validity mask b_i requires at least three preceding positions and nonzero variance. For each source span, IRM retains its top-k (k=2) neighbors as directed edge events, ordered by mean salience w, event count c, and normalized surface text, all descending. Zero-salience candidates can fill remaining slots under this ordering. The weight w is then discarded. Each selected edge stores $\max(c, 1)$ for deletion accounting, while neither count magnitude nor direction enters the retrieval walk.

3.2 The store, symbols only

For each chunk, the store associates a normalized span-and-edge list with the chunk identifier (cid). Raw text and images are retained or linked by the same identifier (Supp. A.3). Every leaf value in the graph records is a string or an integer, with zero latent floats on disk, establishing an ontology-free symbolic memory. Each edge record, an entity pair selected in one chunk with its count, is a *fact*, a triple whose relation slot holds only a single placeholder label. A node's *support* records the chunks that contributed it, one integer per chunk, and its document frequency is the number of those chunks. Supports, co-occurrence counts (the number of chunks contributing a pair) and document frequencies are derived at load. The retrieval graph uses undirected triple-occurrence weights. Stored direction and event-count magnitudes do not enter PPR.

Because records are keyed this way, each chunk's symbolic contributions can be removed exactly. Deleting a chunk removes it, decrements every support and count it contributed, and drops any node or edge that reaches zero. This is exact integer arithmetic with no entangled continuous state, supporting exact and reversible editing of the symbolic memory. Derived IDF values are recomputed after deletion. Store checks verify removal and restoration by deleting a chunk and merging its preserved chunk-level contributions back into the store. Downstream question answering (QA) after editing remains unmeasured (Supp. C.2). Here, "vector-free" describes the persistent symbolic store, and "reader-portable" refers to reuse of its graph with reader-specific preparation.

3.3 Read, a session load then decode-free prefills

Reading starts with reader preparation, which dynamically materializes value representations for reuse across queries. The loaded reader uses its comparison heads to stream the corpus text through a forward pass and regenerate value vectors. These vectors live in a reusable session cache, separate from the persistent symbolic store. A new reader reuses the builder's graph while preparing values in its own semantic space. WRITE and reader preparation use separate forward passes, including when the same model performs both. In the default configuration, one frozen reader handles retrieval and answering without a separate embedder.

For each query, the reader encodes query values and probe spans in separate forward passes. Query value encoding uses a 64-token limit, and the probe pass uses a 128-token limit. For the deployed Qwen reader, the write-time probe also marks query spans, without decoding. A different model needs a separately fitted probe for probe-based seeds.

Across separately encoded contexts, IRM compares tokens in the loaded model's value space. Rotary position embedding (RoPE) rotates queries and keys but not values (Su et al., 2024). IRM uses these unrotated values for cross-context matching. Token-level matching adopts MaxSim late interaction from ColBERT (Khattab & Zaharia, 2020), applied here to un-pooled values regenerated by the frozen reader. By avoiding average pooling, IRM preserves fine-grained semantic resolution, allowing sparse evidence within a passage to contribute strongly to its relevance score. The text-pooling control tests the retrieval benefit of this choice on MuSiQue (Section 4.5). Each query token

contributes its strongest cosine match in the candidate,

$$\mathrm{sim}(Q, \mathcal{C}) \;=\; \frac{1}{|Q|} \sum_{i \in Q} \max_{j \in \mathcal{C}} \cos\big(v_i,\, v_j\big), \tag{3}$$

where Q and $\mathcal{C}$ are the query's and the candidate's token sets, v_i are the query's value vectors from a separate forward pass, and v_j are the candidate's value vectors regenerated at the session load. Text value vectors concatenate values from that reader's four selected KV heads, standardized over corpus tokens, and L2-normalized. Query and cached passage values share the same statistics. For passage seed scores, we weight query-token maxima by selectivity and min–max normalize their sum.

The resulting seeds combine symbolic entity matching for identity and corpus-wide value matching against the session cache for meaning. Entity seeds take the maximum of symbolic top-fact endpoint scores and IDF weights for query-matched entities. Passage seeds use scaled, normalized value scores (Supp. A.3). One personalized PageRank solve (Page et al., 1999), the graph traversal adopted from HippoRAG-2 (Gutiérrez et al., 2025) without iterative reseeding, walks the graph from those seeds to evidence that is not directly named by any seed, producing a shortlist of up to 200 candidates. The shortlist is then re-scored decode-free with the query-likelihood score of UPR (Sachan et al., 2022). Each candidate c is truncated and wrapped in a fixed instruction prefix $T(c)$, then scored by the mean teacher-forced log-likelihood of the query tokens

$$\mathrm{score}(c) \;=\; |q|^{-1} \log P_\theta(q \mid T(c)) \;=\; |q|^{-1} \textstyle\sum_t \log p_\theta(q_t \mid T(c),\, q_{<t}), \tag{4}$$

taken directly from next-token logits under teacher forcing, without generating tokens. After reranking and benchmark-specific document or conversation filtering, up to $K{=}10$ passages go to the frozen reader (Supp. A.3). Value matching seeds graph retrieval, while query likelihood re-scores the shortlist using the same frozen model's full forward computation.

3.4 Multimodal, a pixel-native deferred read

Images are stored as raw pixels and encoded for retrieval only after text retrieval produces a shortlist, avoiding the cost of encoding every corpus image at session load. Retaining the pixels preserves visual details beyond a stored textual description. The frozen reader scores image relevance without generating captions or query expansions. When the query includes an image, the reader compares it with each shortlisted image using Eq. 3 over salient value patches P_q and P_c, with its own image-specific comparison heads for both query and candidate pixels. Each image contributes up to eight patches with the highest value norms, L2-normalized without corpus standardization. Query-likelihood text scores are z-scored over all candidates, while image similarities are z-scored over image-bearing candidates and added to their text scores with weight 1. Candidates without images and text-only queries use only the standardized text scores. The reader answers using the selected passages and their evidence images. Each evidence image appears immediately before its passage's text, and the query image immediately before the question, preserving the image–text associations. This deferred, pixel-native read enables the current model to evaluate query relevance using reader-specific visual representations while retaining the original image evidence for future readers.

4 Experiments

The experiments are designed to test our core hypotheses of answer accuracy supported by structural memory and intrinsic representations, cross-model transferability through representation decoupling, and the computational efficiency of zero-generation memory for lifelong agents. Component controls examine the contributions of topological graph traversal and semantic evidence scoring. The evaluations use static corpora and include exact symbolic deletion/restoration checks.

4.1 Setup

We evaluate whether the memory supports reasoning across documents, long narratives, extended conversations, and visual evidence. These demands are represented by MuSiQue (Trivedi et al., 2022) for multi-hop reasoning over a shared corpus, NarrativeQA (Kočiský et al., 2018) for narrative

Table 1: Architectures of four frozen multimodal models. Qwen3.5-9B is used for answer accuracy and all four models for retention. Attention ratios denote layer counts. GQA denotes grouped-query attention and ViT vision transformer. Model sources are given in Supp. A.2.

Model	Developer	Decoder attention	Vision encoder
Qwen3.5-9B	Alibaba	Gated DeltaNet / full (3:1)	27-layer ViT
Gemma-3-12B-it	Google	Local / global (5:1)	SigLIP
MiMo-VL-7B-RL	Xiaomi	Full attention (GQA)	Native-resolution ViT
Pixtral-12B	Mistral	Full attention (GQA)	Pixtral-ViT (400M)

Table 2: Answer scores (%). Published results are shown alongside re-evaluations and IRM with a common Qwen3.5-9B answer reader. Published readers are Llama-3.3-70B (HippoRAG-2), GPT-3.5-turbo (DRAGON), and Qwen2.5-VL-7B (MuRAG). MuRAG denotes the re-implementation evaluated in MemGallery. Re-evaluation retains each method's prompts and retrieval settings. HippoRAG-2 uses Qwen3.5-9B for OpenIE, recognition filtering and answering. MuSiQue OpenIE uses JSON-constrained decoding (Supp. B.3). Scores use token-overlap F1 except for LoCoMo, whose entries list answerable F1 ($n = 1{,}540$) / adversarial abstention ($n = 446$) / question-weighted overall score, in that order. $\approx$ denotes values derived from rounded published scores. Bold marks IRM scores that exceed the corresponding re-evaluated baseline with the same answer reader.

Benchmark	Compared method	Published (original reader)	Re-evaluated (Qwen3.5-9B)	IRM (ours) (Qwen3.5-9B)
MuSiQue	HippoRAG-2	48.6 (70B)	44.66	37.69
NarrativeQA	HippoRAG-2	25.9 (70B)	23.32	**28.88**
LoCoMo	DRAGON	≈46.4 / 32.5 / 43.3	44.72 / 91.93 / 55.33	**47.45** / 73.54 / 53.31
MemGallery	MuRAG	69.7 (7B)	59.77	**67.02**

comprehension, LoCoMo (Maharana et al., 2024) for long-term dialogue memory including unanswerable questions, and MemGallery (Bei et al., 2026) for questions whose evidence can include images. Table 1 summarizes the model architectures. We use benchmark-specific scorers, including LoCoMo's abstention credit, and HippoRAG-2's released sample and scorer for NarrativeQA (Gutiérrez et al., 2025). A separate cost study uses 50 answerable LoCoMo questions over two complete conversation memories, with common inputs across systems. The reader receives up to $K{=}10$ passages selected from up to $C{=}200$ candidates unless stated otherwise (Supp. A.1 and A.3).

4.2 Answer accuracy

Table 2 compares each benchmark's baseline with IRM using a common answer reader, alongside the baseline's published score. On NarrativeQA, IRM scores 28.88 F1 versus 23.32 for the native-protocol all-Qwen HippoRAG-2 control on the same 293 questions, using up to ten and five passages, respectively. It also exceeds the published 70B-reader system's 25.9 F1, showing that structural memory combined with reader-derived evidence scoring supports competitive narrative question answering with a frozen 9B reader. On LoCoMo, IRM has higher answerable F1 (47.45 versus 44.72), while DRAGON has higher adversarial abstention credit and overall score. On MemGallery, IRM scores 67.02 F1 versus 59.77 for MuRAG with the same answer reader over all 1,711 questions. On MuSiQue, HippoRAG-2 with Qwen3.5-9B for OpenIE, recognition filtering and answering scores 44.66 F1 versus IRM's 37.69 on all 1,000 questions. Published scores use different readers and provide system-level context rather than evidence of a matched-reader advantage.

LoCoMo shows category-specific differences (Table 3). With Qwen3.5-9B answering, IRM exceeds DRAGON on single-hop, multi-hop and open-domain questions, but trails on temporal questions and adversarial abstention. DRAGON's 91.93 abstention credit gives it 55.33 overall versus IRM's 53.31. With the same reader and prompt, IRM exceeds BM25 by 13.92 answerable F1 points, while BM25 has higher adversarial abstention credit (78.70 versus 73.54, Supp. B.3).

Table 3: LoCoMo category scores (%) using official scoring rules, including abstention on adversarial questions. Aggregate scores are question-count-weighted means. DRAGON uses top-5 observations. Its GPT-3.5-turbo scores are published in Table 3 of Maharana et al. (2024), and its 9B scores retain the native protocol with Qwen3.5-9B answering. IRM and BM25 share Qwen3.5-9B, the answer prompt and up to ten evidence turns.

Category	n	IRM (9B)	BM25 (9B)	DRAGON (9B)	DRAGON (GPT-3.5)
Single-hop	841	62.82	48.10	53.01	54.3
Multi-hop	282	40.08	20.14	34.11	36.3
Temporal	321	22.10	15.60	41.63	40.7
Open-domain	96	19.31	5.16	13.70	26.5
Adversarial (abstention)	446	73.54	78.70	91.93	32.5
Answerable mean	1,540	47.45	33.53	44.72	≈46.4
Overall mean	1,986	53.31	43.67	55.33	43.3

Table 4: Each cell reports text / image recall retention, calculated as the reader's recall on the row builder's store divided by its recall on its own store. Values of 1 indicate unchanged recall and values above 1 indicate gains. Text recall uses the full MuSiQue evaluation set (n=1,000, C=200), measured as recall at ten (R@10) before query-likelihood re-scoring. Image recall uses 484 MemGallery queries with mapped gold images, measured at R@10 after text retrieval and patch MaxSim, without answer generation or query-likelihood fusion. Raw scores appear in Supp. C.1.

Built by ↓ / Read by →	Qwen3.5-9B	Gemma-3-12B-it	MiMo-VL-7B-RL	Pixtral-12B
Qwen3.5-9B	1.000 / 1.000	0.948 / 1.000	0.966 / 1.000	0.992 / 1.001
Gemma-3-12B-it	1.004 / 1.000	1.000 / 1.000	1.011 / 1.000	1.034 / 0.999
MiMo-VL-7B-RL	1.002 / 1.000	0.982 / 1.000	1.000 / 1.000	1.021 / 1.001
Pixtral-12B	1.007 / 1.000	0.960 / 1.000	0.977 / 1.000	1.000 / 1.000

4.3 Retention across model swaps

To test our foundational claim of representation decoupling, we keep the stored symbolic graph unchanged across model swaps. Each of the four models serves as both builder and reader, giving sixteen combinations for each modality. Within each column, we keep the reader fixed and vary the model that builds the graph. The retrieval configuration is fixed within each reader column, with representations regenerated by that reader. The text evaluation searches the full 11,656-passage MuSiQue corpus under the protocol in Table 4 and Supp. C. Text recall retention ranges from 0.948 to 1.034, while image recall retention ranges from 0.999 to 1.001. The retained recall provides direct evidence that structural memory remains useful beyond its builder across the tested models. This measures retrieval retention rather than answer-accuracy retention.

4.4 Write and read cost in lifelong agent deployments

We compare IRM and HippoRAG-2 using Qwen3.5-9B for both memory construction and final answering (Table 5). Each system accesses only the question's own conversation. Local models run alone on one RTX PRO 6000 Blackwell Server Edition (96 GB). Both Qwen conditions use sequential Transformers generation for final answering, with their respective retrieval pipelines. HippoRAG-2 uses OpenIE and NV-Embed-v2 (Lee et al., 2025), whereas IRM uses its symbolic graph and reader-derived representations. Under these implementations with the same Qwen model, IRM has lower WRITE time (50.00 versus 3089.60 s) and higher F1 (53.65 versus 47.68), while HippoRAG-2 has lower total READ latency (1.70 versus 2.88 s per query). Query-likelihood re-scoring accounts for 2.48 of IRM's 2.63 s retrieval time.

The Llama condition uses vLLM, while GPT-4o-mini conditions use an application programming interface (API) whose timings include provider and network latency. For IRM with GPT-4o-mini, retrieval and re-scoring remain local with Qwen, and only final answering uses the API. For lifelong agent deployments, generation-free writing introduces a trade-off between construction and query latency. The construction saving can therefore matter most when new memories are written fre-

Table 5: Cost and accuracy on 50 LoCoMo questions from two conversation memories containing 675 and 680 turns. WRITE averages six fresh builds in seconds per memory. READ averages three passes over the questions, and F1 uses the first pass. Retrieve includes refinement and input preparation. Answer includes output processing. Model loading is excluded, and READ measures queries on a ready memory. Baseline indexing is included in WRITE and its index is reused for READ. IRM with GPT-4o-mini combines separately timed local retrieval and API answering. A single model name denotes both WRITE and final answering. Different models are listed as WRITE/READ, and a dash in that column indicates no WRITE LLM. Superscripts identify the answer backend, with HF denoting Hugging Face Transformers. Measurement details appear in Supp. D.1.

			READ (s/query)			
Method	WRITE / READ LLM	WRITE (s)	Retrieve	Answer	Total	F1
BM25	— / GPT-4o-mini$^{\text{API}}$	0.01	0.01	0.50	0.51	14.59
HippoRAG-2	Llama-3.3-70B$^{\text{vLLM}}$	338.15	1.29	3.91	5.19	56.72
HippoRAG-2	GPT-4o-mini$^{\text{API}}$	170.95	0.76	0.92	1.67	47.74
HippoRAG-2	Qwen3.5-9B$^{\text{HF}}$	3089.60	1.45	0.25	1.70	47.68
IRM	Qwen3.5-9B$^{\text{HF}}$	50.00	2.63	0.25	2.88	53.65
IRM	Qwen3.5-9B / GPT-4o-mini$^{\text{API}}$	50.00	2.63	0.50	3.14	45.01

Table 6: Component ablations with Qwen3.5-9B (C=200, K=10). Differences are the component minus its control, in percentage points (pp). Panel A measures gold-passage recall and same-reader F1 on pre-re-score rankings, except the query-likelihood row. Both text-pooling variants disable query-token selectivity weighting to isolate pooling. Panel B measures gold-image recall among images linked to the final ten passages after score fusion, and answer F1. Query pixels remain available in multimodal controls. Caption substitution and omission keep retrieval rankings fixed, so their recall differences are zero (Supp. B.1–B.2).

Component	Control	ΔR@10 (pp)	ΔF1 (pp)
A. MuSiQue text retrieval (n=1,000 for both metrics)			
Graph walk (PPR)	passage value seeds only	+14.19	+5.44
Un-pooled MaxSim	mean-pooled vectors	+5.18	+3.83
Query-likelihood re-score	no re-score	+6.43	+1.68
Attention-salience edges ($\alpha\|v\|$)	co-occurrence (same fan-out)	+0.85	+1.44
Layer/head selection	random layers/heads (5 seeds)	+0.97	+1.72
B. MemGallery multimodal read (n=484 for R@10, n=487 for F1)			
Image re-scoring	no image re-scoring	+51.85	+13.68
Patch MaxSim	mean-pooled patches	−0.32	−1.43
Evidence pixels	dataset captions	0.00	+2.62
Evidence pixels	omitted evidence images	0.00	+0.16

quently, whereas a read-heavy workload can favor lower per-query latency. Timing details are given in Supp. D.1.

4.5 Ablations

The graph walk and query-likelihood re-scoring provide the largest passage-recall gains, while image re-scoring provides the largest gain among the multimodal controls (Table 6). The graph walk adds 14.19 R@10 points over passage value seeds alone. Un-pooled MaxSim adds 5.18 points over mean-pooled vectors. These controls support complementary roles in multi-hop evidence recovery. Topological propagation extends retrieval beyond direct seed matches, while un-pooled matching retains fine-grained token-level relevance. Query-likelihood re-scoring improves R@10 by 6.43 points on the same 200-candidate lists, refining the ranking after candidate retrieval. Attention-salience edges yield observed differences of 0.85 R@10 and 1.44 F1 points over co-occurrence edges with the same nodes and fan-out. This comparison does not establish the superiority of salience-based

connectivity. The joint layer/head selection procedure adds 0.97 R@10 and 1.72 F1 points over the mean of five random-selection runs.

Within the text-retrieved shortlist, image re-scoring combines patch-MaxSim image similarity with query-likelihood text scores. This improves gold-image R@10 by 51.85 points and answer F1 by 13.68 points, with query pixels and the reader fixed across conditions. Mean pooling is a simpler alternative, yielding 1.43 points higher F1 than patch MaxSim and 0.32 points higher recall in this comparison. Providing evidence pixels to the final reader has a smaller effect once retrieval rankings are fixed. Pixels improve F1 over dataset captions by 2.62 points and over omitted evidence images by 0.16 points (Supp. B.2). In this setting, the larger visual gains come from selecting relevant evidence.

5 Analysis

IRM retrieval expands the evidence available beyond lexical matching. On MuSiQue, its pre-re-ranking top ten recover 609 question–gold-passage pairs missed by BM25's top ten, while losing 172 pairs that BM25 retrieves. Of the recovered pairs, 279 lie outside BM25's top 200, so re-ranking that BM25 shortlist alone cannot recover them (BM25 setup in Supp. B.3). This shows complementary evidence recovery. Broad connectivity makes evidence reachable, but useful retrieval still requires selecting the right passages. The largest component contains 99.21% of nodes, and sampled node pairs within that component are 4.12 hops apart on average (Supp. C.2). IDF-weighted seeds and a damped walk guide candidate retrieval, after which query-likelihood re-scoring raises R@10 from 58.06 to 64.49 and complete-gold coverage from 31.0% to 33.9%. The stored links and explicit matching scores make this retrieval process inspectable. The controls establish contributions to evidence recovery, rather than a causal explanation of the reader's reasoning.

6 Limitations and future directions

The symbolic store imposes no rigid ontology, leaving latent entity typing as a way to enrich its semantics. The topology reflects the builder, and retaining at most $k{=}2$ neighbors per entity per chunk can omit useful links, motivating adaptive sparsity. The proper-name span probe remains unmeasured on other languages and scripts. Reader-specific probes are not tested for interchangeability, and comparison-head selection requires evidence-labeled development queries (Supp. A.2). Attention salience provides inspectable connectivity rather than a complete causal explanation. Activation patching could test the selected links.

Answer quality depends on the reader and system configuration, and the common-reader results do not establish uniform accuracy gains. Retention is tested only on 7–12B models using static corpora. Multimodal retrieval remains text-led, so images missed by the text shortlist cannot be recovered by visual re-scoring. Native visual graph traversal is a direction for addressing this limitation.

7 Conclusion

IRM provides an architecture for lifelong agent memory by decoupling structural topology from semantic representation. WRITE constructs a symbolic graph without generating extraction text and retains links to raw evidence. READ separately prepares the current reader's representations and retrieves evidence from text and images. Experiments show the retrieval contributions of graph propagation and token-level matching, and retained recall across the tested model swaps. Construction is faster than the tested extraction-based graph systems in the two-memory study, while per-query latency introduces deployment-dependent trade-offs. Chunk-level accounting supports exact removal and restoration of preserved symbolic contributions. New readers inherit the symbolic store while dynamically regenerating their own representations. IRM thus provides a reader-portable memory foundation for continually evolving multimodal agents.

Reproducibility Statement

Supp. A–D document settings, data, scoring, retention and cost protocols. Code, configurations and saved-output verification are available via the first-page code link.

AI Use Statement

Generative AI assisted conceptual development, experimental design and review, implementation, dataset cleaning and reformatting, result interpretation, literature and policy searches, manuscript drafting and editing, and figure creation, including generated images. Verification included rescoring saved experimental outputs and checking cited sources. The portrait in Figure 1 uses original dataset imagery. The authors are responsible for the final content.

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 *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, 2026. doi: 10.18653/v1/2026.acl-long.1892. URL https://aclanthology.org/2026.acl-long.1892/.

Shijie Chen, Bernal Jiménez Gutiérrez, and Yu Su. Attention in large language models yields efficient zero-shot re-rankers. In *International Conference on Learning Representations (ICLR)*, 2025.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025*, 2025. doi: 10.3233/FAIA251160.

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 GraphRAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020. doi: 10.18653/v1/2020.emnlp-main.574. URL https://aclanthology.org/2020.emnlp-main.574/.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *International Conference on Learning Representations (ICLR)*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

LLM-Core-Team Xiaomi. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025. URL https://arxiv.org/abs/2506.03569.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Mistral AI. Model card for Pixtral-12B-2409, 2024. URL https://huggingface.co/mistralai/Pixtral-12B-2409.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Qwen Team. Qwen3.5-9B: Model card and configuration, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*, 2024.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Song_Deep_Metric_Learning_CVPR_2016_paper.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Qizhi Wang. Democratizing GraphRAG: Linear, CPU-only graph retrieval for multi-hop QA. *arXiv preprint arXiv:2602.23372*, 2025.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations (ICLR)*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3640a1997a4c9571cea9db2c82e1fc35-Abstract-Conference.html.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Chunxu Liu, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. FreeRet: MLLMs as training-free retrievers. In *International Conference on Machine Learning (ICML)*, 2026.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. LinearRAG: Linear graph retrieval augmented generation on large-scale corpora. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=mCtfkypdm6.

SUPPLEMENTARY MATERIAL

A EVALUATION AND MODEL PREPARATION

A.1 EVALUATION SCOPE AND SCORING

Table S1: Benchmark sizes, corpus units, scoring protocols and evidence-label coverage. The final row describes the separate LoCoMo cost study. Gold n counts questions with passage-level evidence labels. For MemGallery, these are official clues mapped to round text, not validated visual-evidence labels.

Benchmark	Questions	Corpus	Scorer	Gold n
MuSiQue	1,000	11,656 passages	Token F1	1,000
NarrativeQA	293	4,072 chunks	Token F1	0
LoCoMo	1,986	5,882 turns	Official score	1,977
MemGallery	1,711	3,962 rounds, 20 scenarios	Official F1	1,527
LoCoMo	50	675/680 turns (2 memories)	Official F1	50

Pools and samples. Text pools use exact full-text deduplication, preserving case and whitespace. MuSiQue uses released HippoRAG-2 passage bodies and questions in file order. Both occurrences of one repeated question are retained (1,000 rows, 999 distinct strings). NarrativeQA and LoCoMo apply document/conversation filtering (Supp. A.3). LoCoMo uses original camera-ready utterances without images or dataset captions. MemGallery retains scenario-specific dialogue rounds and linked images, using the original user and assistant text without appending dataset image captions.

Scoring. MuSiQue averages per-question maximum F1 over answers and aliases after SQuAD-style normalization. Gold-evidence recall is the mean per-question fraction of labeled supporting passages retrieved in the top ten. NarrativeQA takes maximum F1 over two references. LoCoMo uses Porter-stemmed F1, comma-split partial matching for multi-hop questions, the first semicolon-separated reference for open-domain questions, abstention-substring credit for adversarial questions and a question-count-weighted overall score. The nine LoCoMo questions without mapped evidence are excluded only from recall, which uses the top ten turns before conversation filtering. The same scorer applies to all cost-study systems. MemGallery's official F1 uses stemming and the first gold answer.

Published baseline configurations. MuSiQue and NarrativeQA use the Llama-3.3-70B results in Table 2 of Gutiérrez et al. (2025). MuSiQue shares the 11,656 passage bodies, but HippoRAG-2 adds titles and uses five passages with a one-shot reasoning prompt. Published MuSiQue passage R@5 is 43.5 for BM25, 69.7 for NV-Embed-v2 and 74.7 for HippoRAG-2 (Table 3, Gutiérrez et al., 2025). IRM reports 64.49 R@10 on passage bodies without titles. The different cutoffs and passage formatting preclude a matched retrieval comparison. IRM uses body text and Table S3's budgets. NarrativeQA shares question rows and both references, but differs in reader, deduplication, document filtering and evidence budget (5 versus 10 passages). These are system comparisons, not matched-reader controls. MemGallery's published 69.7 F1 comes from its MuRAG re-implementation (Table 4, Bei et al., 2026), using pretrained GME-Qwen2-VL-2B-Instruct embeddings, Qwen2.5-VL-7B reader and top-10 evidence. These baselines do not report gold R@10 under our evidence-label protocol.

A.2 HEAD SELECTION AND ENTITY PROBES

The model sources for Table 1 are Qwen3.5 (Qwen Team, 2026), Gemma 3 (Gemma Team, 2025), MiMo-VL (LLM-Core-Team Xiaomi, 2025) and Pixtral (Mistral AI, 2024). Pixtral uses the `mistral-community/pixtral-12b` conversion and MiMo the `-RL` checkpoint.

Selection procedure. For each eligible (decoder layer, KV head) pair, we score each question's 10 candidate paragraphs using value MaxSim (Eq. 3). We average the AUC over 303 HotpotQA training questions (Yang et al., 2018) and retain the four highest-scoring pairs. This pool shares no question, passage or title with any evaluation corpus. Eligible pairs require a value projection, yielding 32 for Qwen, 384 for Gemma, 288 for MiMo and 320 for Pixtral. Selection scores individual heads using cosine-normalized values (64 query tokens, 480 passage tokens), whereas deployment concatenates four heads, applies corpus standardization and uses 400 passage tokens.

Image-patch heads are selected separately on Stanford Online Products training images (Song et al., 2016), covering 303 product classes in 12 categories, with one query and one positive photo per class (606 images). Each query ranks its same-class positive against other-class photos in the same category. Exact byte/pixel duplicates and near perceptual-hash matches to MemGallery are excluded from selection. Image heads are used only for reading.

Entity probe. At each distinct text-head layer, one probe direction is fitted by logistic regression on 600 passages of the same 303 questions, with rule-generated proper-name labels (capitalized sequences of one to four words with at least three characters, excluding spans whose first word is in a fixed stopword list). Token features use 32-bit floating point (fp32), with a 70/30 token-level training/test split and logistic-regression $C{=}1$. The inference threshold is zero.

Table S2: Text and image head selections, indexed by zero-based (layer, KV head). Entity probes use the distinct text-head layers only.

Model	Text heads	Image heads	Probe layers
Qwen3.5-9B	$(11,1),(7,3)$ $(11,0),(19,0)$ $(17,6),(26,2)$	$(3,1),(3,2)$ $(3,0),(7,3)$ $(3,3),(10,7)$	7,11,19
Gemma-3-12B-it	$(27,3),(23,4)$ $(23,5),(15,6)$	$(3,4),(11,3)$ $(4,1),(5,2)$	17,23,26,27
MiMo-VL-7B-RL	$(12,5),(13,2)$ $(10,3),(11,6)$	$(14,1),(11,7)$ $(4,0),(12,2)$	12,13,15,23
Pixtral-12B	$(19,6),(17,1)$	$(9,6),(7,4)$	10,11,17,19

A.3 PIPELINE AND IMPLEMENTATION SETTINGS

The persistent symbolic store is serialized as positional graph JSON with separate corpus and evidence mappings. Exact prompts, tokenization, normalization, parsing and ordering rules are documented in the released `IMPLEMENTATION.md`. Retrieval and answer generation convert images to RGB and resize them within 384×384 pixels while preserving aspect ratio. Each dialogue round links at most one evidence image. For NarrativeQA and LoCoMo, the final ranking is filtered to the query's document or conversation before selecting up to ten passages, retaining the unfiltered ranking if the filter returns no candidates.

Text QA uses two short-answer demonstrations followed by the retrieved passages and question, with an instruction to return only the answer span. Decoding is greedy, with thinking disabled. Scoring uses a cleaned first-line answer. LoCoMo additionally instructs abstention with "Not mentioned." MemGallery requests concise, dialogue-grounded answers, prioritizing the most recent information. Category-specific formats require image identifiers, yes/no answers or abstention when applicable.

Table S3: Settings used for IRM's main accuracy experiments, grouped into graph construction, retrieval and evidence selection. Diagnostic variants are specified with their experiments.

Parameter	Value	Purpose
Text / image heads	4 / 4	Selected separately on disjoint pools (Supp. A.2)
WRITE edge fan-out k	2	Directed neighbors per entity per chunk
PPR damping	0.5	Graph propagation
Entity seed scale	20	Scale for inverse supporting-chunk frequency
Passage seed weight	0.05	Scale for normalized value-MaxSim scores
Token selectivity weighting	On	Weight query tokens by match selectivity
Re-scoring candidates	200	Passages scored by query likelihood
Reader passages K	10	Maximum passages supplied to answer generation
Symbolic facts linked	5	Highest-scoring entity pairs contributing seeds
Entity surface-key length	4 words	Maximum normalized entity key
WRITE / passage-value limit	400 tokens	Chunk prefill and session value encoding
Query-value / span limits	64 / 128 tokens	Separate value and entity-span passes
Re-score passage limit	1,200 characters	Passage text before prompt wrapping
Re-score prefix limit	800 tokens	Wrapped conditioning prefix
Text QA / MemGallery output limits	30 / 256 tokens	Maximum generated answer length
Image patches	8	Maximum retained patches per image
Image fusion weight	1	Added standardized image score
Image fusion zero threshold	10^{-9}	Population standard deviation at or below threshold
Evidence-image bounds	384×384	Resize within these pixel dimensions

B Component and Answer Evidence

B.1 Text retrieval controls

The query-likelihood R@10 gain is 6.43 points. The matched attention-edge contrast in Table 6 gives a 0.85-point R@10 gain (Qwen build). Its observed F1 difference is 1.44 points. Selectivity weighting adds 1.05 pre-re-score R@10 points on the full MuSiQue evaluation set.

Joint random-selection control. Five fixed seeds (0–4) each sample four distinct eligible attention layers uniformly without replacement, then one KV head uniformly at each layer. Each arm uses its sampled heads for WRITE and READ and fits layer probes on the same external 600-passage pool. Random-selection R@10 is 57.09 ± 1.58 (seed mean $\pm$ standard deviation (SD)), versus selected 58.06. The selected configuration has four heads on three layers, whereas these controls have four heads on four layers. The comparison measures the joint selection procedure, not an isolated probe or head effect. Random-selection F1 is 34.29$\pm$0.99 across seeds.

B.2 Multimodal controls and category results

Ranking controls. For the controls in Table 6, the text-led shortlist and query-likelihood scores remain fixed. Patch pooling also retains the selected image heads and fusion rule. For evidence-image recall, duplicate images linked to the final ten passages count once. Later passages do not fill missing image slots, and the query image is not retrieved evidence. This differs from the image-only top-ten ranking in Supp. C.1. Text/image subsets refer to query modality, not supporting-evidence modality. Full IRM reaches image R@10 83.85 under these controls.

Evidence controls. On the image-query subset with fixed rankings, F1 is 82.08 with evidence pixels, 79.46 with dataset captions and 81.92 when evidence images are omitted. Caption and omission controls (Table 6) retain the caption-stripped text corpus. Passages retain turn text without prepended session dates and appear in retrieval order. The omission comparison concerns evidence pixels given pixel-informed retrieval and query pixels, not an entirely text-only pipeline.

Table S4: MemGallery token F1 by question group over all 1,711 questions, using scenario-specific memories.

Question group	n	F1
Extraction	862	81.34
Reasoning	503	42.29
Management	346	67.28

B.3 Matched answer controls and evidence recovery

Recovery beyond lexical retrieval. BM25 uses `rank_bm25` default BM25Okapi parameters and lowercase alphanumeric tokenization on the same MuSiQue passage bodies and question rows. It achieves 41.39 R@10 and 12.6% complete-gold coverage. Complete-gold coverage requires every gold passage for a question.

Matched candidate-generation control. On the same 1,000 questions and 11,656 passage bodies, we pass BM25's top 200 candidates through the unchanged IRM UPR stage and Qwen3.5-9B reader (K=10), retaining the prompts, truncation, decoding and scorer. Table S5 reports the scores. This compares candidate-generation systems and does not isolate attention-derived edges from IRM's other retrieval components.

Table S5: MuSiQue candidate-generation control on the same 1,000 questions. BM25+UPR and IRM share the Qwen3.5-9B answer reader, prompt, scorer, UPR and evidence budgets. Scores are percentages.

System	K	Gold R@10	F1
BM25+UPR	10	56.78	33.49
IRM	10	64.49	37.69

Reference-protocol common-reader evaluation. The reference controls in Table 2 retain each method's prompts, retrieval units and scorer. Generation settings follow the released protocol, except for JSON-constrained OpenIE on MuSiQue and reconstructed MuRAG server defaults as detailed below. The NarrativeQA, LoCoMo and MemGallery controls cover 293, 1,986 and 1,711 questions, respectively. The MuSiQue control scores 44.66 F1 on 1,000 questions using Qwen3.5-9B for OpenIE, recognition filtering and answering. NarrativeQA uses Qwen3.5-9B for OpenIE, recognition filtering and answering. Both HippoRAG-2 controls use $K = 5$, NV-Embed-v2 and no UPR. MuSiQue retains titles, native prompts, greedy decoding and the 2,048-token answer limit, served by vLLM. All 1,000 outputs are scored, including 176 that reach the token limit and 137 for which the native parser uses the full response because the answer marker is absent. DRAGON retains released observations, date-prefixed DRAGON encoding, top-5 retrieval, the native temporal and adversarial prompts, and a 32-token answer limit over all 1,986 LoCoMo questions. To accept the native system-only request, the Qwen chat template removes only its user-message presence check, preserving the system role, prompt content and non-thinking assistant prefix. MemGallery's MuRAG implementation retains GME-Qwen2-VL-2B cosine top-10 retrieval and native text/image prompts over 1,711 questions. Only its final reader changes to Qwen3.5-9B. The MemGallery API request leaves the output limit to the server. Its historical server flags are not published, so the remeasurement records reconstructed no-override reference defaults separately. These are common-reader system comparisons, retaining each method's design. They are distinct from the shared-prompt candidate-generation control.

C Cross-model Reuse and Store Verification

C.1 Cross-reader recall

Table S6 reports the absolute recall underlying Table 4. Within each reader column, only the builder's symbolic graph changes. The reader's own heads, probes, text/image representations, corpus and retrieval budget remain fixed. Text recall uses all 1,000 MuSiQue questions.

Raw text and pixel grid. Text uses macro gold-passage R@10. Unlike the evidence-image control in Supp. B.2, the pixel grid ranks unique image files attached to the graph-retrieved top-200 passages using the current reader's image heads and evaluates an image-only top ten, under Table 4's protocol.

Table S6: Absolute retrieval R@10 (%), shown as Text / Image for the same builder–reader pairs as Table 4. Image scores use 484 queries with mapped gold images and an image-only top ten, rather than images attached to the final ten passages. Each reader's diagonal score is the normalization denominator.

Built by ↓ / Read by →	Qwen3.5-9B	Gemma-3-12B-it	MiMo-VL-7B-RL	Pixtral-12B
Qwen3.5-9B	58.06 / 94.99	59.17 / 84.43	56.72 / 93.26	57.52 / 67.27
Gemma-3-12B-it	58.28 / 94.99	62.42 / 84.43	59.39 / 93.26	59.97 / 67.17
MiMo-VL-7B-RL	58.17 / 94.99	61.31 / 84.43	58.73 / 93.26	59.17 / 67.29
Pixtral-12B	58.46 / 94.99	59.94 / 84.43	57.38 / 93.26	57.98 / 67.23

C.2 Symbolic persistence and editing

The full MuSiQue store has 56,978 entity nodes and 244,032 stored directed edge records, corresponding to 200,860 unique undirected pairs. Its largest component contains 99.21% of nodes. The symbolic store occupies 13.1 MB on disk, excluding reader-specific token values.

Store edits. On the full 11,656-chunk store, exact deletion/restoration passes 50 gold-evidence cases and 200 additional edits. Each deleted state also matches re-aggregating the remaining stored chunk records, including derived IDF. Disjoint half-store merge/split is exact. These tests cover the symbolic store and derived IDF, not live cache invalidation, physical removal of linked image files, or model unlearning. The implementation scans the stored records and rebuilds IDF.

D Cost Measurement

D.1 Measurement protocol

A fixed hash order selects two LoCoMo conversations and 25 answerable questions per conversation, allocated proportionally by category. Timestamped speaker turns exclude images, captions and added summaries. Gold evidence is used only for evaluation. WRITE includes extraction, indexing and persistence but excludes warm-up. READ measures queries on a ready memory. For IRM/Qwen, retrieval ends after final-answer prompt tokenization. For HippoRAG-2/Qwen, it ends immediately before dispatch to the local Qwen server. API/vLLM conditions likewise end retrieval at final-QA request dispatch. HippoRAG-2's recognition filter is included in retrieval. Model loading is excluded from WRITE and READ.

IRM uses the accuracy experiments' frozen selection and bfloat16 (bf16) Qwen reader. BM25 uses whitespace tokenization and the IRM/GPT answer configuration. HippoRAG-2 uses NV-Embed-v2 with GPT-4o-mini or Llama-3.3-70B-Instruct with 4-bit weights and 16-bit activations (W4A16), served by vLLM. The HippoRAG-2/Qwen condition uses Qwen for OpenIE and the same final-answer template as IRM/Qwen. Its Answer timer includes client cleaning, scoring and per-query persistence. HippoRAG-2 accesses Qwen through a local HTTP server, while IRM runs the reader in process. The 150 READ measurements use the two conversation indices. WRITE is measured across six builds. Timing-receipt persistence itself is outside the measured interval. Queries run sequentially without competing measurements. Query-embedding and response-cache reuse are disabled for timing, while the prepared passage-value cache is retained across queries. Prefix/provider

caching retains its defaults. Across six builds, IRM WRITE is 50.004 ± 0.344 s (mean $\pm$ sample SD), shared by both reader conditions. HippoRAG-2/Qwen WRITE is 3089.596 ± 31.761 s. Its 150-query READ is 1.703 ± 0.436 s/query, with first-pass F1 47.68. Execution details are in `TABLE5.md`. WRITE includes chunk prefill, but incremental capture during an existing prefill is unmeasured. Query-likelihood re-scoring costs 2.48 s/query for both IRM conditions.