

MEMORA: A Memory-Enhanced Multimodal Committee Reranking Agent for Reasoning-Intensive Retrieval

Mohamed Mahmoud¹, Mostafa Farouk Senussi¹, Mahmoud Abdalla¹,
Mahmoud SalahEldin Kasem¹, Abdelrahman Abdallah², Hyun Soo Kang¹

¹Chungbuk National University ²University of Innsbruck
hskang@cbnu.ac.kr

Abstract

Reasoning-intensive retrieval is fundamentally challenged when queries demand deep semantic inference and arrive in multimodal form. Existing systems treat each query in isolation, rely on monolithic rerankers, and lack retrievers trained for cross-modal reasoning. We introduce **MEMORA**¹, a **Memory-Enhanced Multimodal Orchestrated Reranking Agent** built around three contributions: (1) **Collaborative Factor-Committee Reranking**, which decomposes queries into intent factors, extracts factor-specific evidence into a structured matrix, and synthesizes judgments via a memory-conditioned meta-verdict model; (2) an **Episodic Memory Module** that tracks factor outcomes across queries and dynamically adjusts priorities at inference time; and (3) a **Reasoning-Aware Dense Retriever** fine-tuned with chain-of-thought augmentation on 20k+ reasoning-intensive pairs. Evaluated on MM-BRIGHT across 29 domains, MEMORA achieves nDCG@10 of **41.5** (text-to-text) and **38.9** (multimodal query-to-text), outperforming the strongest agentic text-only baseline by **+8.3** points and the strongest agentic multimodal baseline by **+7.9** points.

1 Introduction

Retrieval systems are increasingly expected to handle queries that demand genuine reasoning rather than surface-level matching. This challenge is compounded in real-world settings, where users combine natural language with images — attaching circuit diagrams to technical questions, screenshots of error logs to debugging requests, or scientific figures to research queries. In these settings, identifying the relevant document requires understanding what the image depicts, how it relates to the textual question, and which documents in the corpus

collectively address both dimensions. Yet most retrieval systems are designed for text-only queries, and those that support multimodal input lack the reasoning capacity to handle this level of cross-modal inference.

This gap is starkly exposed by MM-BRIGHT (Abdallah et al., 2026), the first benchmark for reasoning-intensive multimodal retrieval. Despite significant advances in vision-language models, the best multimodal encoder achieves only 27.6 nDCG@10 on the multimodal-to-text task — *lower* than the best text-only retriever (32.2), meaning that adding visual information actively *hurts* performance. This counterintuitive result reveals three compounding failure modes that existing approaches have not addressed jointly: (a) **Monolithic reranking**: existing rerankers apply a single, uniform reasoning pass over all candidates, regardless of how heterogeneous the query’s underlying information needs are; (b) **Stateless retrieval**: systems treat every query independently, discarding accumulated knowledge about which retrieval strategies succeed or fail within a domain; and (c) **Weak multimodal grounding**: standard dense encoders are not trained to handle the abstract, cross-modal reasoning that multimodal queries demand, producing embeddings where visual signals add noise rather than signal. Prior work addresses at most one of these failures in isolation (Sun et al., 2023; Weller et al., 2025; Wang et al., 2023; Gao et al., 2023; Zhang et al., 2024; Jiang et al., 2024).

We introduce **MEMORA** (Memory-Enhanced Multimodal Orchestrated Reranking Agent), a unified agentic framework that addresses all three failure modes jointly. As illustrated in Figure 1, MEMORA operates through four tightly integrated stages: (1) a **Reasoning-Aware Dense Retriever** encodes queries enriched with chain-of-thought traces to produce a high-quality initial candidate pool; (2) an **Episodic Memory Module** retrieves

¹Code and data are available: <https://github.com/HarnessLab/MEMORA>.

similar past queries from domain memory and uses accumulated factor outcome experience to decompose the current query into prioritized intent factors; (3) a **Collaborative Factor Committee** issues factor-specific subqueries, extracts targeted evidence per candidate, and organizes judgments into a bipartite **factor-document evidence matrix** that a memory-conditioned meta-verdict model consults to produce the final ranking; and (4) an **Adaptive Controller** dynamically decides whether to expand retrieval, invoke the judge, or stop based on ranking margin and evidence support, making MEMORA responsive to query difficulty rather than fixed in its computation.

Our contributions are as follows:

- We identify three compounding failure modes in reasoning-intensive multimodal retrieval — monolithic reranking, stateless retrieval, and weak multimodal grounding — and propose **MEMORA**, a unified agentic framework that addresses all three jointly (§3).
- We introduce **Collaborative Factor-Committee Reranking**, in which query intent is decomposed into factors, evidence is extracted per factor and organized into a bipartite factor-document graph, and a memory-conditioned meta-verdict model produces the final ranking (§3).
- We propose an **Episodic Memory Module** that accumulates domain-specific factor outcome experience across queries, dynamically adjusting factor priorities at inference time without retraining, and train a **Reasoning-Aware Dense Retriever** on 20k+ reasoning-intensive Stack Exchange pairs with CoT annotation and hard negatives (§3.2).
- The full MEMORA system achieves **41.5** and **38.9** nDCG@10 on MM-BRIGHT Tasks 1 and 2, outperforming the strongest agentic text-only baseline by **+8.3** points and the strongest agentic multimodal baseline by **+7.9** points (§5.1).

2 Related Work

Retrieval benchmarks and models. Early benchmarks such as BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) focus on surface-level relevance, while BRIGHT (Su et al., 2025)

and RAR-b (Xiao et al., 2024) introduced reasoning as a core requirement, exposing sharp performance drops. MM-BRIGHT (Abdallah et al., 2026) extends this to the multimodal domain, with concurrent benchmarks MRMR (Zhang et al., 2025) and ARK (Lin et al., 2026) confirming that visual and spatial reasoning remain persistent bottlenecks. Dense retrievers such as E5 (Wang et al., 2024), GritLM (Muennighoff et al., 2024), and SFR-Embedding (Meng et al., 2024) perform well under standard conditions but struggle under reasoning pressure; ReasonIR (Shao et al., 2025) and DiVeR (Sun et al., 2025) partially address this through reasoning-aware training. Multimodal retrievers including VLM2Vec (Jiang et al., 2024), GME (Zhang et al., 2024), and Nomic-Vision (Nussbaum et al., 2024) extend dense retrieval to vision-language settings, yet their best result of 27.6 nDCG@10 on MM-BRIGHT Task 2 still falls below the best text-only baseline, confirming that visual signals add noise without explicit reasoning.

Reranking, query expansion, and memory. LLM-based rerankers (Zhuang et al., 2023; Pradeep et al., 2023) substantially outperform dense retrieval alone; RankGPT (Sun et al., 2023) pioneered listwise reranking, while Rank1 (Weller et al., 2025) and RankR1 (Zhuang et al., 2025) further incorporate chain-of-thought reasoning and reinforcement learning. Query expansion methods (Gao et al., 2023; Wang et al., 2023; Lei et al., 2025) and agentic systems (Trivedi et al., 2023; Chen et al., 2026) improve recall through LLM-generated subqueries and interleaved reasoning. Memory-augmented approaches draw on episodic experience for LLMs (Fountas et al., 2024), case-based RAG (Wiratunga et al., 2024), and multimodal episode indexing (Xie et al., 2024). Despite this progress, no prior work addresses all three failure modes jointly: existing rerankers are monolithic, retrievers are stateless, and no system unifies episodic factor memory with committee-based reranking under multimodal conditions.

3 Method

MEMORA is a memory-conditioned retrieval agent that treats each query $q = (t, x)$ — text t and optional image x — as a composition of latent intent factors. Given a corpus $\mathcal{D} = \{d_j\}_{j=1}^N$, the goal is to produce a ranking $\pi(q) = (d_{(1)}, \dots, d_{(K)})$ that maximizes relevance to the full user intent.

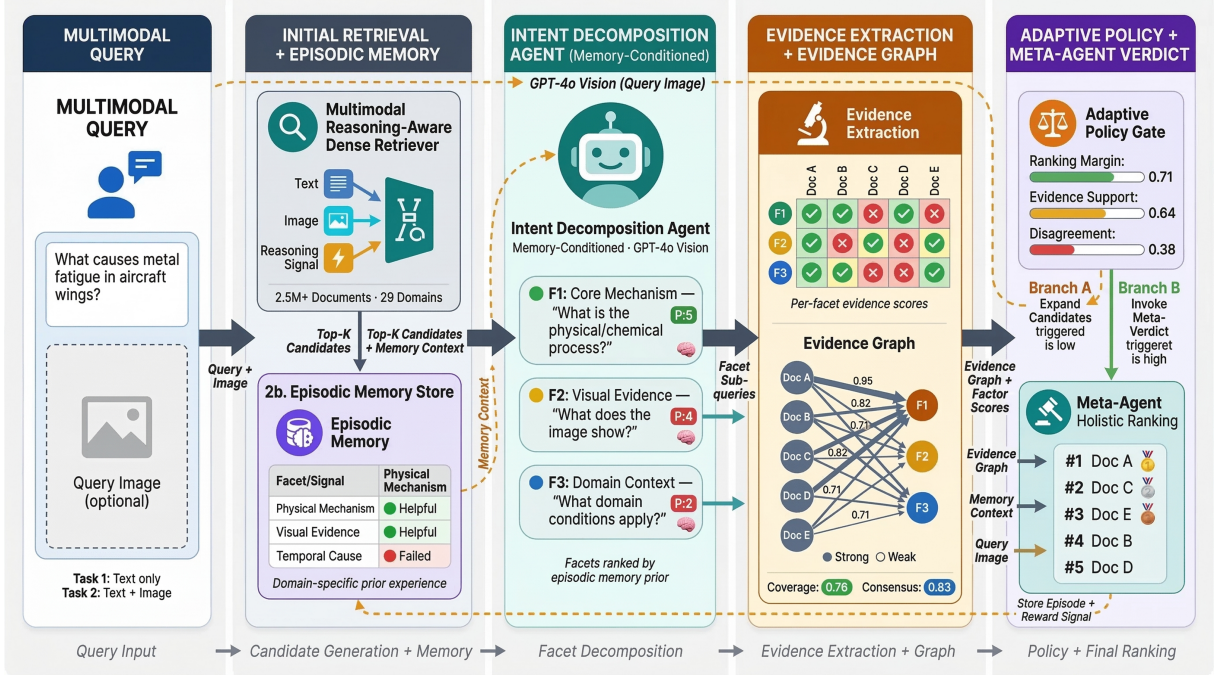


Figure 1: Overview of MEMORA. Given a multimodal query (t, x) , the Episodic Memory Module retrieves domain-specific prior experience to guide factor decomposition and candidate expansion. The Multimodal Reasoning-Aware Dense Retriever produces an initial candidate pool, which is expanded via factor-specific subqueries. A factor evidence extractor then builds a bipartite factor-document evidence graph, whose structured signals feed into a memory-conditioned meta-verdict model for final ranking. An Adaptive Controller decides at each step whether to expand retrieval, invoke the judge, or stop.

Unlike fixed prompt pipelines, MEMORA persistently stores which factors helped or failed in previous queries and uses this signal to guide both retrieval and ranking.

3.1 Overview

The pipeline proceeds in four stages: (1) a reasoning-aware dense retriever produces an initial candidate pool; (2) domain memory episodes guide decomposition of the query into intent factors; (3) factor-specific subqueries expand the pool and a factor-document evidence graph is constructed; (4) a factor-aware scorer ranks candidates, with a memory-aware listwise judge invoked when needed.

3.2 Reasoning-Aware Dense Retriever

Standard dense retrievers are not trained for the deep semantic inference that reasoning-intensive queries demand. We therefore train a dedicated retriever on a purpose-built corpus of reasoning-intensive pairs as the backbone of MEMORA.

Architecture. We adopt a bi-encoder built on **Qwen-Embedding**. The query encoder ϕ_q produces $\mathbf{q} = \phi_q(q) \in \mathbb{R}^d$; document embeddings

$\mathbf{d}_j = \phi_d(d_j)$ are pre-computed offline. Retrieval scores are:

$$s_j^{(0)} = \mathbf{q}^\top \mathbf{d}_j. \quad (1)$$

Training Data. We construct **20,000+ reasoning-intensive question-answer pairs** from Stack Exchange, where each question is augmented with a GPT-4o chain-of-thought annotation articulating the latent retrieval intent. Hard negatives are mined via an initial BM25 pass.

Reasoning Signal. The query encoder receives the concatenation $[r; q]$, where r is the CoT annotation, teaching it to embed latent retrieval intent rather than surface features. At inference time the encoder operates on the raw query alone, having internalized this behavior during training.

Training. We fine-tune with in-batch and hard negatives via InfoNCE:

$$\mathcal{L} = -\log \frac{\exp(\mathbf{q}^\top \mathbf{d}^+ / \tau)}{\exp(\mathbf{q}^\top \mathbf{d}^+ / \tau) + \sum_j \exp(\mathbf{q}^\top \mathbf{d}_j^- / \tau)}, \quad (2)$$

where $\mathbf{d}^+$ is the positive document, $\mathbf{d}_j^-$ are hard negatives, and τ is a temperature hyperparameter.

3.3 Initial Retrieval

The top- K_0 documents scored by R form the initial candidate set $\mathcal{C}^{(0)}$. For multimodal queries, R consumes both t and x ; for text-only queries, x is omitted.

3.4 Domain Memory

For each domain c , MEMORA maintains an episodic memory $\mathcal{M}_c = \{e_i\}_{i=1}^{T_c}$. Each memory entry stores the previous query, its factor graph, evidence snippets, factor scores, factor outcomes, and final reward:

$$e_i = (q_i, \mathcal{G}_i, \mathcal{S}_i, \Delta_i, r_i). \quad (3)$$

Here, $\mathcal{G}_i$ is the stored factor graph, $\mathcal{S}_i$ contains factor scores per document, and Δ_i summarizes which factors helped or failed.

Given a new query q , we retrieve the most similar episodes from the same domain using a lightweight hybrid similarity:

$$\text{sim}(q, e_i) = \lambda_J J(T(q), T(q_i)) + (1 - \lambda_J) \rho(q, q_i), \quad (4)$$

where $T(\cdot)$ extracts content tokens, $J(\cdot, \cdot)$ is Jaccard similarity, and $\rho(\cdot, \cdot)$ is a string-level similarity score. The top- K_m episodes form the retrieved memory set $\mathcal{M}(q) \subseteq \mathcal{M}_c$.

From $\mathcal{M}(q)$, we compute two types of priors. The first is a document prior

$$m_j = \frac{1}{|\mathcal{M}(q)|} \sum_{e_i \in \mathcal{M}(q)} \mathbb{I}[d_j \in \text{TopDocs}(e_i)], \quad (5)$$

which tracks documents that repeatedly appeared in related episodes; this prior is only beneficial when combined with factor outcome signals, as it can otherwise introduce frequency bias toward common documents over genuinely relevant ones (see §5.3). The second is a factor prior: for each factor label f_k , memory returns a helpfulness prior h_k and a failure prior ℓ_k , estimated from previous outcomes stored in Δ_i .

3.5 Memory-Conditioned Factor Decomposition

Instead of treating the query as a single unit, MEMORA decomposes it into a small set of intent factors

$$F(q) = \{f_1, \dots, f_K\}. \quad (6)$$

Each factor f_k contains a factor name, a focused factor question, a retrieval-oriented subquery, and

a priority score. The number of factors K is itself adaptive: it is set by the controller based on the number of memory hits $|\mathcal{M}(q)|$ and whether an image is present, within the range $[K_{\min}, K_{\max}]$. The factor planner is conditioned on the query and the retrieved memory summary, so the decomposition is informed by both the current input and past domain experience. For multimodal queries, the factor planner is additionally conditioned on a visual description of x produced by a vision-language model (GPT-4o), enabling factors to explicitly reference visual content.

Let $p_k \in \{1, \dots, 5\}$ denote the planner’s initial priority for factor f_k . We then adjust this priority using memory:

$$\tilde{p}_k = \text{clip}\left(p_k + \mathbb{I}[h_k > \tau_+] - \mathbb{I}[\ell_k > \tau_-], 1, 5\right). \quad (7)$$

This simple rule increases the weight of factors that have historically helped and suppresses factors that repeatedly over-supported distractors.

3.6 Factor-Guided Retrieval Expansion

The initial pool $\mathcal{C}^{(0)}$ may miss documents needed by one specific factor. To improve recall, MEMORA may issue factor-specific retrieval calls. For factor f_k , let z_k denote its subquery. The retriever returns

$$s_{j,k}^{\text{exp}} = R(z_k, d_j). \quad (8)$$

We normalize all retrieval scores and combine the initial retrieval, factor expansion, and memory prior into a single candidate score:

$$\hat{s}_j = \max\left(\bar{s}_j^{(0)}, \max_k \bar{s}_{j,k}^{\text{exp}}\right) + \lambda_m m_j. \quad (9)$$

The refined candidate set is then

$$\mathcal{C} = \text{TopK}_{K_c}(\{\hat{s}_j\}_{j=1}^N). \quad (10)$$

3.7 Factor Evidence Extraction

For each candidate document $d_j \in \mathcal{C}$ and each factor f_k , an evidence extractor returns a factor relevance score $a_{j,k} \in [0, 10]$ and a short evidence snippet $u_{j,k}$:

$$(a_{j,k}, u_{j,k}) = E(q, f_k, d_j). \quad (11)$$

Scores are normalized within each factor:

$$\bar{a}_{j,k} = \frac{a_{j,k}}{\max_{j'} a_{j',k} + \epsilon}. \quad (12)$$

This step yields factor-specific evidence rather than a single opaque relevance judgment.

3.8 Factor Evidence Graph

The core representation in MEMORA is a bipartite graph $\mathcal{G} = (\mathcal{V}_d, \mathcal{V}_f, \mathcal{E})$, where document nodes $\mathcal{V}_d = \{d_j\}$ connect to factor nodes $\mathcal{V}_f = \{f_k\}$. An edge $(d_j, f_k) \in \mathcal{E}$ carries two attributes: the normalized factor score $\bar{a}_{j,k}$ and the extracted snippet $u_{j,k}$. The edge weight $w_{j,k} = \bar{a}_{j,k}$ encodes how strongly document d_j satisfies factor f_k , enabling the meta-verdict judge to consult structured per-factor support rather than a single opaque relevance score.

3.9 Factor-Aware Document Scoring

A good document should satisfy multiple factors simultaneously. We therefore combine four signals: average factor support, factor coverage, factor consensus, and retrieval prior. For document d_j , we define

$$u_j = \frac{1}{K} \sum_{k=1}^K \bar{a}_{j,k}, \quad (13)$$

$$c_j = \frac{1}{K} \sum_{k=1}^K \mathbb{I}[\bar{a}_{j,k} \geq \tau], \quad (14)$$

$$v_j = \min_k \bar{a}_{j,k}, \quad (15)$$

where u_j measures overall support, c_j measures how many factors are covered, and v_j penalizes documents that fail any single critical factor. The final factor-based score is

$$s_j = \lambda_u u_j + \lambda_c c_j + \lambda_v v_j + \lambda_r \hat{s}_j, \quad (16)$$

where the weights $\lambda_u, \lambda_c, \lambda_v, \lambda_r$ balance structural evidence against retrieval confidence.

3.10 Adaptive Controller

MEMORA does not execute the same action sequence for every query. Instead, an adaptive controller decides whether to expand retrieval, invoke the final judge, or stop. At step t , the controller observes a compact state vector $z_t \in \mathbb{R}^5$: normalized ranking margin, memory hit ratio, mean evidence support, Spearman correlation between retriever-based and factor-based rankings, and a binary image indicator. The action is selected as

$$a_t = \arg \max_{a \in \mathcal{A}} g_a(z_t), \quad g_a(z_t) = w_a^\top z_t + b_a, \quad (17)$$

where $\mathcal{A} = \{\text{expand, judge, stop}\}$. The weights w_a and biases b_a are set by cross-validation on a held-out development split. The linear form is

intentional: it is interpretable, adds negligible inference overhead, and can be replaced with a learned policy in future work.

3.11 Memory-Aware Final Ranking and Memory Update

When ambiguity remains after factor scoring, MEMORA invokes a listwise meta-verdict judge that receives the query, optional image, top candidate documents, the factor evidence graph, and the retrieved memory summary. The judge consults the factor graph rather than raw documents alone, keeping the final ranking grounded in explicit structured evidence.

After ranking, MEMORA computes which factors were useful and writes this signal back to memory. For factor f_k , let μ_k^+ and μ_k^- denote the average factor scores over relevant documents $\mathcal{D}^+$ and non-relevant candidates $\mathcal{D}^-$:

$$\mu_k^\pm = \frac{1}{|\mathcal{D}^\pm|} \sum_{d_j \in \mathcal{D}^\pm} \bar{a}_{j,k}. \quad (18)$$

The helpfulness and failure signals are then:

$$\begin{aligned} \delta_k^+ &= \max(0, \mu_k^+ - \mu_k^-), \\ \delta_k^- &= \max(0, \mu_k^- - \mu_k^+). \end{aligned} \quad (19)$$

These are written back to domain memory as the helpfulness and failure priors h_k and ℓ_k for future queries, closing the memory loop illustrated in Figure 1.

4 Experiments

4.1 Dataset

We evaluate on MM-BRIGHT (Abdallah et al., 2026), the first reasoning-intensive multimodal retrieval benchmark, comprising **2,803 queries** across **29 technical domains** in four categories (STEM & Life Sciences, Software & Technical Systems, Social Sciences & Humanities, and Applied Domains) over a corpus of **2.5M documents** from Stack Exchange. We evaluate on **Task 1** (text-to-text) and **Task 2** (multimodal query-to-text, where queries include diagrams, charts, and screenshots), using **nDCG@10** as the primary metric.

To prevent training-test contamination, we constructed MEMORA-R’s training set by excluding all Stack Exchange threads whose post IDs appear in the MM-BRIGHT document corpus, as well as any threads posted after the benchmark’s data collection cutoff. The resulting training split is temporally and ID-disjoint from both the retrieval corpus and the query set used for evaluation.

Table 1: Task 1 (text-to-text) nDCG@10 per domain across 29 MM-BRIGHT domains. Each retriever is paired with SelfRAG, Search-o1, and MEMORA as reranking layers.

Domain	BM25	BM25+ SelfRAG	BM25+ Search-o1	BM25+ Memora	E5	E5+ SelfRAG	E5+ Search-o1	E5+ Memora	DiVeR	DiVeR+ SelfRAG	DiVeR+ Search-o1	DiVeR+ Memora
<i>STEM & Life Sciences</i>												
Acad	9.1	11.1	10.5	15.2	24.6	26.6	30.1	35.7	31.6	34.0	37.2	44.9
Bio	4.8	6.4	7.5	10.4	18.2	21.4	27.1	36.1	27.7	32.5	40.0	54.3
Chem	7.8	8.8	10.2	11.8	25.4	26.6	28.8	32.1	33.8	35.2	32.8	41.8
Phys	4.0	5.3	6.2	10.2	14.6	15.1	15.9	17.1	19.9	20.0	20.2	20.6
Math	2.0	3.1	4.9	10.5	18.1	19.3	21.4	24.6	34.2	35.0	32.5	38.8
Earth	4.7	5.8	7.5	11.1	29.1	31.0	34.4	39.7	34.6	36.7	40.4	46.2
BioAc	8.5	10.2	15.1	18.1	20.5	21.8	24.0	27.5	27.4	28.4	30.1	32.9
BioInf	5.3	6.4	8.4	10.5	18.9	19.1	19.4	19.8	32.7	32.7	31.7	29.3
Med	12.6	13.2	11.2	15.4	34.7	36.1	38.7	42.7	36.1	37.6	40.2	44.4
<i>Software & Technical Systems</i>												
Apple	1.0	2.0	5.1	7.2	24.8	24.9	25.1	25.3	22.8	23.0	21.3	23.8
Ubuntu	17.4	15.1	20.0	22.0	29.0	29.8	31.2	33.3	44.6	45.3	43.5	48.5
BTC	3.3	4.3	5.1	7.0	32.4	33.1	34.4	36.4	30.4	31.1	31.4	34.4
Crypto	0.8	3.1	6.2	8.7	10.2	10.7	11.5	12.8	24.3	24.3	21.3	18.9
GIS	1.3	2.1	4.5	10.5	25.0	25.0	25.0	25.0	31.7	31.7	30.7	28.4
QC	3.8	4.2	6.2	12.5	11.4	11.7	12.4	13.3	13.0	13.1	11.2	13.5
Robot	5.0	6.0	9.7	11.8	20.2	20.8	21.8	23.5	33.4	33.4	30.5	33.6
Sales	3.9	4.8	3.1	12.1	25.7	27.7	31.2	36.7	38.6	40.9	41.9	51.2
<i>Social Sciences & Humanities</i>												
Econ	4.0	5.0	6.1	9.8	27.7	29.2	31.7	35.8	31.9	33.4	31.1	40.4
Psych	5.3	5.4	7.9	10.7	22.8	24.4	27.3	31.8	29.3	31.2	31.5	39.7
Phil	4.1	5.3	7.7	15.2	21.3	22.8	25.4	29.4	17.4	19.0	20.8	26.2
Law	6.2	7.5	10.1	14.9	50.2	52.4	56.3	62.4	52.3	54.5	54.5	64.7
Christ	30.3	20.1	25.1	35.1	29.9	30.7	32.0	34.1	35.2	37.4	40.2	47.3
Islam	9.8	10.2	12.8	17.2	27.5	28.1	29.2	31.0	34.9	35.2	31.8	36.7
<i>Applied Domains</i>												
Aviat	1.1	2.2	7.1	6.8	21.9	22.6	24.0	26.1	29.5	30.1	30.3	33.1
Game	36.0	32.1	40.2	41.0	42.1	44.5	48.9	55.7	50.4	54.9	50.0	75.6
PM	18.0	17.0	15.1	20.7	31.0	31.0	31.0	31.0	40.4	35.4	30.4	37.8
Sustain	11.2	12.0	15.1	19.2	30.6	32.4	35.7	40.7	34.8	30.7	35.1	45.4
Travel	22.2	23.4	25.1	24.8	24.8	25.6	27.1	29.4	38.3	40.9	41.6	53.0
Quant	2.3	5.1	7.2	10.1	22.5	24.3	27.5	32.4	23.2	25.0	28.2	33.2
Avg.	8.5	8.9	11.1	14.8	25.3	26.5	28.6	31.8	32.2	33.2	33.2	39.3

4.2 Experimental Setup

The retriever was trained on $4 \times \text{H100 80GB GPUs}$ with distributed data-parallel training (bf16). All LLM calls — factor decomposition, evidence extraction, and the meta-verdict judge — use **GPT-4o**, with GPT-4o Vision generating image descriptions for multimodal queries. Document embeddings are cached and reused across runs. Unless stated otherwise, we use $K_0=50$, $K_c=30$, $K=10$, $K_{\min}=2$, $K_{\max}=5$ (default $N=3$), $K_m=4$, $\lambda_m=0.1$, $(\lambda_u, \lambda_c, \lambda_v, \lambda_r)=(0.4, 0.3, 0.2, 0.1)$, and LLM temperature 0.2 / 0.0 for evidence extraction and the judge.

4.3 Baselines

Retriever Baselines. For Task 1: **BM25** (sparse), **E5-Mistral-7B** (Wang et al., 2024) (dense), and **DiVeR** (Sun et al., 2025) (reasoning-aware, state-of-the-art on BRIGHT). For Task 2 we additionally evaluate multimodal retrievers **BGE-VL** (Zhou et al., 2024), **CLIP** (Radford et al., 2021), and

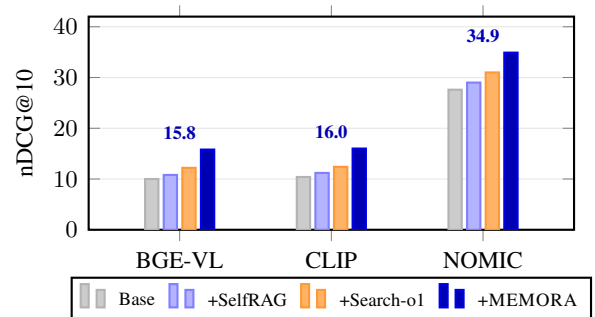


Figure 2: Task 2 average nDCG@10 across 29 MM-BRIGHT domains. MEMORA consistently outperforms SelfRAG and Search-o1 across all three multimodal base retrievers. Full per-domain results in Appendix A.

Nomic-Vision (Nussbaum et al., 2024) (best published Task 2 result: 27.6 nDCG@10).

Agentic Reranking Baselines. Each retriever is paired with **SelfRAG** (Asai et al., 2023), which selectively retrieves and critiques passages, and **Search-o1** (Li et al., 2025), which combines iter-

Table 2: Component analysis of MEMORA. All configurations use DiVeR (Task 1) and Nomic-Vision (Task 2) as base retrievers.

Configuration	nDCG@10 T1	nDCG@10 T2
Base retriever only	32.2	27.6
+ Factor decomposition only	33.5	28.9
+ Factor expansion	34.6	30.4
+ Evidence graph + factor scoring	36.8	32.2
+ Episodic memory	36.4	33.1
+ Meta-verdict judge	37.9	33.9
Full MEMORA (+ adaptive controller)	39.3	34.9
w/o episodic memory	35.5	30.3
w/o factor expansion	34.4	33.4
w/o meta-verdict judge	35.5	31.5
w/o adaptive controller	36.7	32.1

ative search with chain-of-thought reasoning, to isolate MEMORA’s agentic contribution from the underlying retriever.

4.4 Main Results

Task 1: Text-to-text retrieval. Table 1 reports nDCG@10 across all 29 domains. MEMORA consistently improves over every base retriever: +6.3 points over BM25 (8.5 $\rightarrow$ 14.8), +6.5 over E5-Mistral-7B (25.3 $\rightarrow$ 31.8), and +7.1 over DiVeR (32.2 $\rightarrow$ **39.3**), the strongest text-only result, outperforming the next best agentic reranker Search-o1 by +6.1 points. Gains are largest where domain-specific factor memory provides most leverage: Gaming (+25.2), Travel (+14.7), and Law (+12.4). The only exception is Crypto (DiVeR+MEMORA: 18.9 vs. DiVeR: 24.3), where cryptocurrency terminology causes the factor planner to generate overlapping subqueries with low discriminative power.

Task 2: Multimodal query-to-text retrieval. Table 8 reports Task 2 results. CLIP (10.4) and BGE-VL (10.0) perform far below text-only DiVeR (32.2), confirming that standard vision-language encoders fail under reasoning pressure. MEMORA substantially recovers this gap: paired with Nomic-Vision (27.6), it reaches **34.9** (+7.3 over base, +3.9 over Search-o1); CLIP+MEMORA (16.0) nearly doubles CLIP alone. Gains are strongest in STEM domains requiring cross-modal reasoning (Mathematics: 34.0 $\rightarrow$ 42.5; Bioinformatics: 33.8 $\rightarrow$ 42.2) and GIS, where visual-spatial factor decomposition proves particularly effective (Nomic: 25.8 $\rightarrow$ 32.2).

Component analysis. Table 2 traces each component’s contribution (DiVeR / Nomic-Vision as base retrievers). Factor decomposition alone yields

Table 3: Standalone retriever comparison on MM-BRIGHT (nDCG@10). MEMORA-R is evaluated *without* any agentic reranking.

Model	Type	T1	Δ T1	T2	Δ T2
BM25	Sparse	8.5	-23.7	-	-
Contriever	Dense	20.1	-12.1	-	-
E5-Mistral-7B	Dense	25.3	-6.9	-	-
GritLM-7B	Dense	25.3	-6.9	-	-
SFR-Embedding	Dense	26.9	-5.3	-	-
ReasonIR	Reasoning	28.6	-3.6	-	-
DiVeR	Reasoning	32.2	0.0	-	-
CLIP	Vision-Lang	-	-	10.4	-17.2
BGE-VL	Vision-Lang	-	-	10.0	-17.6
GME	VLM	-	-	22.0	-5.6
Nomic-Vision	VLM	-	-	27.6	0.0
MEMORA-R (ours)	Ours	34.1	+1.9	33.4	+5.8

modest gains (+1.3 T1/T2); factor-guided expansion adds +1.1/+1.5 by surfacing domain-specific documents missed by the base retriever. The largest single gain comes from the evidence graph and factor scoring (+2.2 T1, +1.8 T2). Episodic memory contributes more on T2 (+0.9) than T1 (+0.4), reflecting its value for cross-modal factor prioritization. The meta-verdict judge (+1.5/+0.8) and adaptive controller (+1.4/+1.0) complete the system. Removing any single component degrades performance, with episodic memory and the meta-verdict judge being the most critical.

Table 4: Training ablation of MEMORA-R. Each row adds one ingredient on top of the previous. Δ is cumulative gain over the contrastive FT baseline. nDCG@10.

Configuration	T1	Δ T1	T2	Δ T2
Base Qwen-Embedding (no fine-tuning)	24.3	-	25.1	-
+ Contrastive FT (InfoNCE, baseline)	28.1	0.0	28.9	0.0
+ Hard negative mining	31.1	+3.0	30.7	+1.8
+ Reasoning trace augmentation	32.8	+4.7	31.8	+2.9
Full MEMORA-R	34.1	+6.0	33.4	+4.5

5 Additional Studies

5.1 Standalone Retriever Evaluation

Table 3 evaluates MEMORA-R before any agentic reranking. It achieves **34.1** T1 (+1.9 over DiVeR) and **33.4** T2 (+5.8 over Nomic-Vision) — notably, as a text-only encoder at inference time, it outperforms all dedicated multimodal retrievers, confirming that CoT-augmented training captures latent multimodal intent without direct visual encoding. Using MEMORA-R as the first-stage retriever, the full system achieves **41.5** T1 and **38.9** T2, outperforming the best agentic baselines by +8.3 and +7.9 points respectively (Table 7, Appendix B).

Table 5: nDCG@10 as a function of episodic memory depth. Results with MEMORA-R as base retriever.

Episodes stored	T1 nDCG@10	T2 nDCG@10
0 (no memory)	35.5	30.3
1–10	36.1	32.1
11–25	37.2	33.5
26–50	37.4	33.6
>50	41.5	38.9

Table 6: Memory signal decomposition. Each row activates a different subset of the episodic memory signals; all other MEMORA components are held fixed. nDCG@10 with MEMORA-R.

Memory signals used	T1	T2
No memory	35.5	30.3
Document prior only (m_j)	33.4	28.1
Factor prior only (h_k, ℓ_k)	34.6	30.4
Memory summary $\rightarrow$ decomposition only	36.5	31.9
Memory summary $\rightarrow$ meta-judge only	35.7	32.0
All memory signals (full MEMORA)	41.5	38.9

5.2 Retriever Training Ablation

Table 4 isolates each training ingredient. The zero-shot Qwen-Embedding baseline (24.3 T1, 25.1 T2) already shows reasonable capability; contrastive fine-tuning on Stack Exchange pairs brings the first substantial gain (+3.8 T1/T2). Hard negative mining adds +3.0/+1.8 by forcing the encoder to distinguish semantically similar but non-relevant documents, and reasoning trace augmentation contributes a further +1.7/+1.1 by conditioning on CoT annotations that articulate latent retrieval intent. Each ingredient provides an additive benefit, yielding +6.0 T1 and +4.5 T2 over the contrastive FT baseline for the full MEMORA-R.

5.3 Memory Effectiveness Analysis

To evaluate memory depth without leakage, we split each domain’s queries chronologically: the first 80% form the warm-up pool whose episodes are written before test time, and the remaining 20% are held out as test queries. Table 5 reports performance as a function of the number of warm-up episodes available per domain, showing monotonic improvement from 35.5/30.3 (no memory) to **41.5/38.9** (>50 episodes). The largest gains accrue on T2 (+8.6), confirming that episodic memory is especially valuable for cross-modal queries where factor priorities are harder to set without prior domain experience. Domains with dense query populations — Law, Gaming, Travel, and Biology —

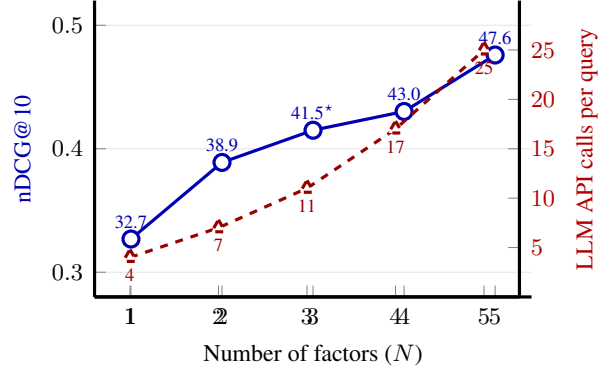


Figure 3: T1 nDCG@10 (blue, left axis) and LLM API calls per query (red dashed, right axis) as a function of the number of intent factors N . Performance improves monotonically but LLM calls grow super-linearly, as each additional factor requires evidence extraction and subquery retrieval calls across all candidates.

benefit most, with individual gains exceeding +10 nDCG@10 at full memory depth.

Table 6 decomposes memory signals. The document prior alone (m_j) slightly *hurts* performance (T1: 33.4 vs. 35.5; T2: 28.1 vs. 30.3), as it boosts documents that appeared frequently in past episodes regardless of whether those retrievals were successful, introducing a frequency bias toward common documents over genuinely relevant ones. The factor prior (h_k, ℓ_k) consistently helps by conditioning on outcome signals rather than appearance counts, and routing the memory summary to factor decomposition provides the single largest individual gain (36.5 T1, 31.9 T2) — confirming that prior experience is most valuable when shaping *which* factors are generated rather than adjusting downstream scoring alone.

5.4 Analysis

Effect of intent factor count. Figure 3 plots T1 nDCG@10 and LLM API calls per query as a function of the number of intent factors N . Performance improves monotonically from 32.7 at $N = 1$ to 47.6 at $N = 5$, confirming that richer intent decomposition consistently benefits retrieval quality. However, LLM API calls grow super-linearly with N — from 4 calls at $N = 1$ to 25 at $N = 5$ — because each additional factor incurs evidence extraction and subquery retrieval across the full candidate pool. We therefore adopt $N = 3$ as the default, which recovers **87.2%** of the $N = 5$ gain at **44%** of the inference cost, representing the best nDCG@10-to-cost operating point on this curve. Sensitivity analyses of candidate pool size K_0 and

failure mode analysis are provided in Appendix C.

6 Conclusion

We presented MEMORA, a memory-enhanced multimodal reranking agent addressing three compounding failures in reasoning-intensive retrieval: monolithic reranking, stateless retrieval, and weak multimodal grounding. MEMORA achieves **41.5** and **38.9** nDCG@10 on MM-BRIGHT Tasks 1 and 2, outperforming the strongest agentic text-only baseline by +8.3 points and the strongest agentic multimodal baseline by +7.9 points across all 29 domains. Future work will explore robust factor generation under lexically ambiguous queries and cross-domain memory transfer for low-resource settings.

Limitations

MEMORA relies on GPT-4o for factor decomposition, evidence extraction, and the meta-verdict judge, which introduces API cost and latency that may limit deployment in low-resource settings. The episodic memory requires a warm-up period of >50 queries per domain before yielding its largest gains, making it less effective for rare or newly-added domains. The Reasoning-Aware Dense Retriever is trained on Stack Exchange data and may not generalize equally to domains with substantially different writing styles. Finally, the adaptive controller uses a fixed linear decision rule that, while interpretable, may be suboptimal compared to a learned policy.

Acknowledgments

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2026-RS-2020-II201462, 50%), and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (RS-2023-NR076833).

References

Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and Hyun-Soo Kang.

2026. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2601.09562*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Jimmy Lin, Akari Asai, and Victor Zhong. 2026. Agentir: Reasoning-aware retrieval for deep research agents. *arXiv preprint arXiv:2603.04384*.

Zafeirios Fountas, Martin A Benfeghou, Adnan Omerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. 2024. Human-inspired episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Yibin Lei, Tao Shen, and Andrew Yates. 2025. Thinkqe: Query expansion via an evolving thinking process. *arXiv preprint arXiv:2506.09260*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438.

Yijie Lin, Guofeng Ding, Haochen Zhou, Haobin Li, Mouxiang Yang, and Xi Peng. 2026. Ark: A dual-axis multimodal retrieval benchmark along reasoning and knowledge. *arXiv preprint arXiv:2602.09839*.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfembedding-mistral: enhance text retrieval with transfer learning. *Salesforce AI Research Blog*, 3(6).

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

- Zach Nussbaum, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed vision: Expanding the latent space. *arXiv preprint arXiv:2406.18587*.
- Ronak Pradeep, Sahel Sharifmoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and 1 others. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.
- Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary Siegel, Michael Tang, and 1 others. 2025. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. 2025:48941–48991.
- Duolin Sun, Meixiu Long, Dan Yang, Junjie Wang, Yecheng Luo, Yue Shen, Jian Wang, Hualei Zhou, Chunxiao Guo, Peng Wei, and 1 others. 2025. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14918–14937.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.
- Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.
- Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*.
- Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruwan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.
- Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*.
- Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. 2024. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313*.
- Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025. Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. *arXiv preprint arXiv:2510.09510*.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.
- Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.
- Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2308–2313.
- Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

A Per-Domain Task 2 Results

Figure 2 summarizes Task 2 results (full per-domain breakdown in Table 8, Appendix A) across all 29 MM-BRIGHT domains. Each multimodal base retriever (BGE-VL, CLIP, Nomic-Vision) is paired with SelfRAG, Search-o1, and MEMORA as reranking layers. MEMORA consistently improves over both agentic baselines across all domains and retrievers, with the largest gains in Law (Nomic: 47.6 $\rightarrow$ 59.5), Gaming (Nomic: 43.1 $\rightarrow$ 53.9), and Mathematics (Nomic: 34.0 $\rightarrow$ 42.5). The only domain where gains are modest across all retrievers is Quantum Computing, where the sparse query population limits the episodic memory’s ability to accumulate reliable factor priors.

Table 7: Full end-to-end system performance. nDCG@10 averaged over 29 domains. T1 = text-to-text; T2 = multimodal query-to-text.

Configuration	T1 nDCG@10	T2 nDCG@10
DiVeR (best published text retriever)	32.2	–
Nomic-Vision (best published MM retriever)	–	27.6
DiVeR + SelfRAG	33.2	–
DiVeR + Search-o1	33.2	–
Nomic + SelfRAG	–	29.0
Nomic + Search-o1	–	31.0
DiVeR + MEMORA	39.3	–
Nomic + MEMORA	–	34.9
MEMORA-R + MEMORA	41.5	38.9

B Full System Results with Trained Backbone

Table 7 reports end-to-end performance when MEMORA-R is used as the first-stage retriever instead of a third-party backbone. The full system achieves **41.5** T1 and **38.9** T2, adding +2.2 T1 and +4.0 T2 over the best third-party pairings (DiVeR+MEMORA: 39.3; Nomic+MEMORA: 34.9), confirming that the reasoning-aware retriever and agentic reranking framework are synergistic.

C Analysis

Effect of candidate pool size. Figure 4 shows T1 nDCG@10 as a function of K_0 . Performance increases from 38.0 ($K_0 = 25$) to 41.5 ($K_0 = 50$) and 43.0 ($K_0 = 75$), with diminishing returns beyond the default $K_0 = 50$. The +3.5 gain from 25 to 50 reflects a higher-quality candidate pool for the factor committee, while the smaller +1.5 gain from 50 to 75 suggests marginal documents at larger pool sizes add noise to evidence extraction. We adopt $K_0 = 50$ as the default.

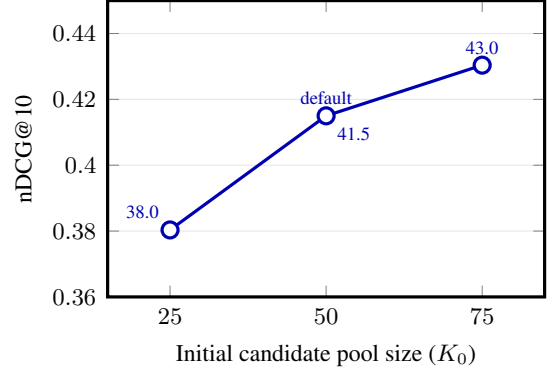


Figure 4: T1 nDCG@10 as a function of the initial candidate pool size K_0 . Performance improves consistently with larger pools, with diminishing returns beyond $K_0 = 50$ (default). Results with MEMORA-R + full MEMORA on MM-BRIGHT.

Failure modes. Two recurring patterns emerge from per-domain results. First, high-ambiguity domains — Crypto (T1: 18.9 vs. base 24.3) and GIS (T1: 28.4 vs. 31.7) — see MEMORA underperform the base retriever, as the factor planner generates semantically overlapping subqueries that introduce noise rather than diversity. Second, low-resource domains with fewer than 10 stored episodes (e.g., Quantum Computing, Philosophy) show modest memory gains, confirming that episodic memory requires sufficient query volume to learn reliable factor priors. Both failure modes point to natural future directions: more robust factor generation under ambiguous queries, and cross-domain memory transfer to bootstrap low-resource domains.

Domain	BGE-VL	BGE-VL+ SelfRAG	BGE-VL+ Search-o1	BGE-VL+ Memora	CLIP	CLIP+ SelfRAG	CLIP+ Search-o1	CLIP+ Memora	NOMIC	NOMIC+ SelfRAG	NOMIC+ Search-o1	NOMIC+ Memora
<i>STEM & Life Sciences</i>												
Acad	4.2	4.5	5.1	9.2	4.8	5.2	5.8	8.0	22.6	23.7	25.3	30.2
Bio	5.7	6.2	7.0	10.4	14.8	16.0	17.8	21.5	26.9	28.2	30.1	33.6
Chem	10.8	11.7	13.2	16.0	9.6	10.4	11.5	13.9	30.6	32.1	34.3	38.2
Phys	6.8	7.3	8.3	11.1	6.1	6.6	7.3	10.8	17.2	18.1	19.3	21.5
Math	13.1	14.1	16.0	19.4	17.9	19.3	21.5	26.0	34.0	35.7	38.1	42.5
Earth	10.1	10.9	12.3	14.9	10.9	11.8	13.1	15.8	30.1	31.6	33.7	37.6
BioAc	13.3	14.4	16.2	20.7	11.4	12.3	13.7	16.5	23.4	24.6	26.2	29.2
BioInf	11.6	12.5	14.2	18.2	9.4	10.2	11.3	13.6	33.8	35.5	37.9	42.2
Med	12.6	13.6	15.4	18.6	9.8	10.6	11.8	14.2	33.9	35.6	38.0	42.4
<i>Software & Technical Systems</i>												
Apple	7.2	7.8	8.8	12.7	12.3	13.3	14.8	17.8	28.7	30.1	32.1	35.9
Ubuntu	11.6	12.5	14.2	17.2	5.5	5.9	6.6	11.0	34.3	36.0	38.4	42.9
BTC	8.9	9.6	10.9	15.2	8.3	9.0	10.0	12.0	22.7	23.8	25.4	28.4
Crypto	11.3	12.2	13.8	16.7	14.8	16.0	17.8	21.5	22.4	23.5	25.1	28.0
QC	4.5	4.9	5.5	10.7	2.6	2.8	3.1	8.8	12.1	12.7	13.6	18.1
Robot	16.1	17.4	19.6	23.8	10.6	11.4	12.7	15.4	30.3	31.8	33.9	37.9
Sales	14.2	15.3	17.3	21.0	2.3	2.5	2.8	9.3	26.2	27.5	29.3	32.8
GIS	13.8	14.9	16.8	20.4	13.1	14.1	15.7	19.0	25.8	27.1	28.9	32.2
<i>Social Sciences & Humanities</i>												
Econ	9.5	10.3	11.6	14.1	6.0	6.5	7.2	10.7	21.1	22.2	23.6	28.4
Psych	6.4	6.9	7.8	10.5	8.7	9.4	10.4	12.6	23.9	25.1	26.8	29.9
Phil	2.4	2.6	2.9	7.6	5.4	5.8	6.5	10.8	21.7	22.8	24.3	27.1
Law	10.2	11.0	12.4	15.1	19.7	21.3	23.6	28.6	47.6	50.0	53.3	59.5
Christ	8.9	9.6	10.9	13.2	15.0	16.2	18.0	21.8	30.9	32.4	34.6	38.6
Islam	12.0	13.0	14.6	17.8	10.7	11.6	12.8	15.5	28.9	30.3	32.4	36.1
<i>Applied Domains</i>												
Aviat	9.6	10.4	11.7	15.2	15.4	16.6	18.5	22.3	24.1	25.3	27.0	32.1
Game	17.5	18.9	21.3	25.9	19.1	20.6	22.9	27.7	43.1	45.3	48.3	53.9
PM	8.6	9.3	10.5	13.7	8.9	9.6	10.7	12.9	27.6	29.0	30.9	34.5
Sustain	10.1	10.9	12.3	17.9	9.0	9.7	10.8	13.0	24.7	25.9	27.7	31.9
Travel	10.1	10.9	12.3	18.9	16.1	17.4	19.3	23.3	36.7	38.5	41.1	45.9
Quant	8.1	8.7	9.9	12.0	2.1	2.3	2.5	10.0	16.2	17.0	18.1	20.2
Average	10.0	10.8	12.2	15.8	10.4	11.2	12.4	16.0	27.6	29.0	31.0	34.9

Table 8: Task 2 (multimodal query-to-text) nDCG@10 per domain across 29 MM-BRIGHT domains. Queries include images such as diagrams, charts, and screenshots paired with natural language questions. Each multimodal retriever is paired with SelfRAG, Search-o1, and MEMORA as reranking layers.