# GraMRAG: Orchestrating Multi-Agent Multi-Step Reasoning via Graph Memory with Reinforcement Learning

**Anonymous ACL submission**


## Abstract

Although existing multi-agent Retrieval-Augmented Generation (RAG) systems have demonstrated promise on complex multimodal reasoning tasks, they remain fundamentally limited in reasoning depth and memory structure, suffering from inadequate retrieval and state blindness when answering knowledge-intensive questions. To address these limitations, we propose GraMRAG, a graph memory-guided multi-agent RAG framework that integrates a dynamic multimodal memory graph to enable stable, multi-step multimodal reasoning. We introduce a vision-text bridged reasoning paradigm that unifies multi-scale entity cropping with a ReAct-style visual toolchain, enhancing the long-horizon cross-modal reasoning capability. We further construct a multimodal memory graph that formalizes agent reasoning as a dynamic directed acyclic graph (DAG), explicitly modeling action-observation dependencies to mitigate state blindness and suppress redundant retrieval. Moreover, we propose Topology-Aware Policy Optimization (TAPO) that leverages graph topology for critical path identification and targeted node pruning, enabling fine-grained credit assignment across multi-step reasoning trajectories. Extensive experiments on challenging multimodal benchmarks demonstrate that our approach consistently outperforms existing baselines and achieves state-of-the-art performance on complex long-horizon reasoning tasks.


## 1 Introduction

*"Never memorize something that you can look up."*
— Albert Einstein

Retrieval-Augmented Generation (RAG) extends the knowledge boundaries of Multimodal Large Language Models (MLLMs) by retrieving relevant information from external knowledge bases (Fan et al., 2026; Hu et al., 2026). Recent advances in MLLMs have convincingly demonstrated that multimodal agentic RAG systems are capable of efficient retrieval and reasoning over large-scale corpora comprising interleaved textual and visual content (Zhang et al., 2026; Guan et al., 2025). MLLM-based multi-agent architectures, leveraging specialized collaboration among components, exhibit superior performance over single-agent frameworks on complex multimodal RAG tasks (Zhou et al., 2026; Chang et al., 2025; Xu et al., 2025b).

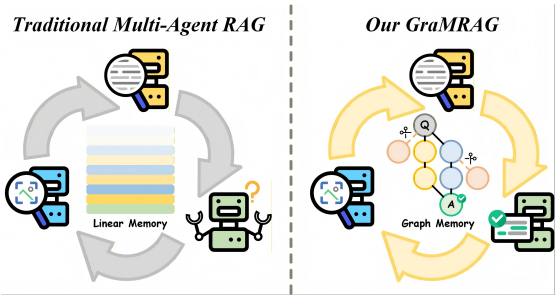


Figure 1: Comparison between our GraMRAG and existing multi-agent RAG systems.

However, existing multi-agent RAG systems face three key limitations. First, the average number of retrieval rounds in current systems remains limited, predisposing agents to prematurely terminate exploration and settle for partially available evidence on complex tasks (Jiang et al., 2025; Wu et al., 2025a; Team et al., 2025). This fundamentally impedes the transfer of the long-horizon planning capability from text-based deep research to multimodal scenarios, rendering these systems inadequate for tens of iterative reasoning steps demanded by real-world deep research. Second, prevailing linear or iterative summarization-based memory structures (Yao et al., 2022; Zhou et al., 2025; Xu et al., 2025a; Wu et al., 2025b), lacking explicit modeling of agent reasoning states, are prone to context expands and state blindness as interaction rounds accumulate, resulting in repetitive queries and redundant retrieval in multi-hop tasks. Third, conventional outcome-based sparse

rewards coarsely propagate terminal-state signals across entire trajectories (Huang et al., 2026; Zhou et al., 2025), causing effective retrieval steps to be erroneously penalized in negative samples while redundant operations are undeservedly rewarded in positive ones, preventing fine-grained credit assignment across multi-step reasoning.

To address the mentioned limitations, we propose GraMRAG, a novel graph memory-guided multi-agent RAG framework. First, to bridge the long-horizon reasoning gap, we introduce a vision-text bridged reasoning paradigm that orchestrates multi-scale entity cropping with a ReAct-style visual toolchain, seamlessly transferring the iterative planning capability of text-only deep research models into the multimodal domain to enable long-horizon cross-modal reasoning trajectories. Second, to overcome state blindness and context expansion in linear memory paradigms, we construct a multimodal memory graph that formalizes the reasoning process as a dynamic directed acyclic graph (DAG), explicitly encoding agent actions and multimodal observations with their temporal and logical dependencies to mitigate redundant retrieval and state blindness. Third, to tackle the credit assignment problem from coarse terminal rewards, we introduce a Topology-Aware Policy Optimization (TAPO) algorithm that leverages the graph topology to identify critical reasoning paths and perform targeted node pruning, precisely masking false positive and false negative nodes during policy updates to enable fine-grained credit assignment across multi-step reasoning.

Our contributions are summarized as follows:

- We introduce a vision-text bridged reasoning paradigm that unifies multi-scale entity cropping with a ReAct-style visual toolchain, enabling long-horizon cross-modal reasoning.
- We construct a multimodal memory graph that formalizes agent reasoning as a dynamic DAG to systematically mitigate state blindness and redundant retrieval.
- We develop a TAPO algorithm leveraging graph topology for critical path identification and node pruning, enabling fine-grained credit assignment across multi-step reasoning.
- We conduct extensive experiments on challenging multimodal benchmarks to evaluate our approach, demonstrating state-of-the-art performance over existing baselines.

## 2 Related Work

### 2.1 Multi-Agent Systems in Multimodal RAG

Multi-agent systems have demonstrated promise on complex multimodal reasoning tasks (Dong et al., 2026; Li et al., 2026; Xinjie et al., 2025; Zhang et al., 2025a). By deploying multiple specialized agents that each focus on distinct facets of a problem, these architectures enable collaborative problem-solving that surpasses the capabilities of single-agent frameworks (Ni and Ye, 2026; Sinha et al., 2026; Siingh et al., 2025). However, existing multi-agent RAG systems conduct insufficient retrieval rounds, causing agents to prematurely terminate exploration on complex tasks (Geng et al., 2025; Narayan et al., 2025). Such shallow reasoning trajectories are fundamentally insufficient for demanding deep research scenarios that require tens of iterative reasoning steps. To address this challenge, we introduce a vision-text bridged reasoning paradigm integrating multi-scale entity cropping with a visual toolchain, using image descriptions as a cross-modal bridge to enable long-horizon planning in the multimodal domain.

### 2.2 Memory Structures of Agent Systems

The memory architectures of existing agent systems can broadly be categorized into two paradigms (Zhang et al., 2025b; Ye et al., 2025): the history accumulation paradigm, exemplified by ReAct (Yao et al., 2022), which linearly concatenates all interaction records, and the iterative summarization paradigm, exemplified by Mem1 (Zhou et al., 2025), which controls context length by compressing observations. However, as the number of interaction rounds grows, both paradigms cause agents to progressively lose track of historical retrieval paths, a phenomenon we term state blindness, which in turn induces redundant queries and repetitive retrieval in multi-hop tasks. To address this issue, we construct a multimodal memory graph that formalizes the reasoning process as a dynamic DAG, explicitly encoding the temporal and logical dependencies between actions and observations, thereby effectively mitigating state blindness and redundant retrieval.

## 3 Method

### 3.1 Vision-Text Bridged Reasoning Paradigm

**Multi-Scale Visual Entity Grounding.** Given an input image $I$, a user query $q$, and a visual in-

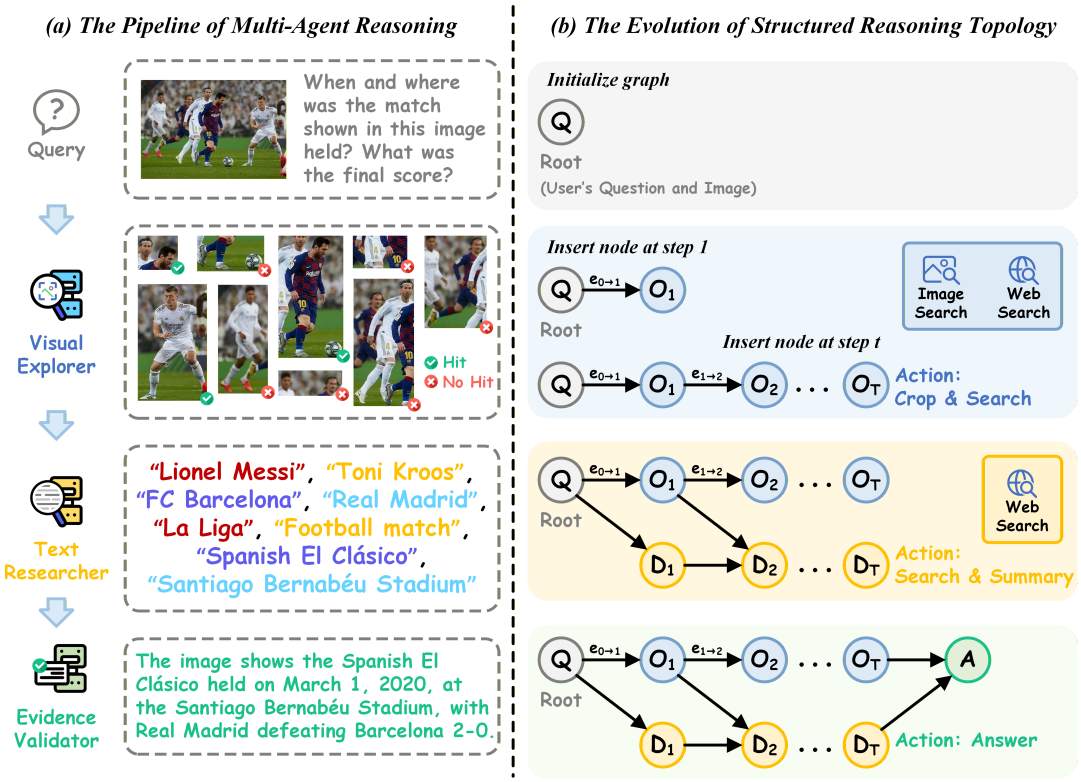


Figure 2: Overview of the GraMRAG framework.

duction prompt $p_v$, the Visual Explorer performs hierarchical region proposal generation at each reasoning step $t$. Rather than submitting the full image as a single retrieval unit, which is prone to failure under background noise interference, the Visual Explorer simultaneously generates coarse-level contextual regions and fine-grained entity-level patches, forming a candidate bounding-box set $\mathcal{B}_t = \{b_t^1, b_t^2, \ldots, b_t^{N_t}\}$, where each $b_t^i$ represents a region crop at a distinct spatial scale or entity granularity, and $N_t$ denotes the number of proposals at step $t$. This candidate set induces a visual retrieval action $A_t^v = \mathrm{VSearch}(\mathcal{B}_t)$, which is processed sequentially by a two-stage visual toolchain: an image search module that matches each cropped region against a web-scale index and returns candidate URLs and a webpage retrieval module that fetches page content in structured markdown format. The observation obtained at step $t$ is denoted:

$$O_t^v = \mathrm{VisChain}(A_t^v) \tag{1}$$

Upon completing the current step, the Visual Explorer applies the induction prompt $p_v$ again to generate the next search context, producing the visual reasoning trajectory:

$$\begin{aligned}\tau_{\text{vis}} = \{I,\ q,\ p_v,\ R_1,\ A_1^v,\ O_1^v,\\ \ldots,\ p_v,\ R_{T_v},\ A_{T_v}^v,\ O_{T_v}^v\}\end{aligned} \tag{2}$$

where $R_t$ denotes the chain-of-thought reasoning generated prior to action $A_t^v$, and $T_v$ is the total number of visual retrieval steps.

Following each visual retrieval action, an independent Hit/No Hit Judge evaluates the observation $O_t^v$ obtained at the current step, conditioned on the original image $I$ and query $q$, outputs a binary acceptance signal:

$$s_t = \mathrm{Judge}(I,\ q,\ O_t^v) \in \{0, 1\} \tag{3}$$

When $s_t = 1$ (Hit), the retrieval is considered successful, meaning the retrieved content is semantically consistent with the query entity, and $O_t^v$ is added into the accumulated valid evidence pool:

$$\mathcal{E}^v = \{O_t^v \mid s_t = 1,\ t = 1,\ \ldots,\ T_v\} \tag{4}$$

When $s_t = 0$ (No Hit), the retrieval is considered unsuccessful, triggering the Visual Explorer to re-crop at a finer spatial granularity or to reformulate the search query before initiating a new retrieval attempt. The visual exploration phase terminates when $|\mathcal{E}^v|$ reaches a predefined sufficiency threshold or when the total step count reaches $T_v$. Decoupling the Hit/No Hit judgment from the main agents and assigning it to an independent judge achieves a clean separation between exploration strategy and evidence evaluation, enhancing the robustness of the multi-scale iterative retrieval process.

**Modality Bridging and Textual Reasoning.** To transfer the evidence accumulated during the visual exploration phase into the textual reasoning domain, we introduce a modality bridging step. A captioning module first generates a comprehensive textual description $D$ of the input image $I$, covering key entities, spatial relationships, and scene-level semantics. The visual trajectory $\tau_{\text{vis}}$ is then transformed into a bridged context: $I$ is replaced by $D$, all visual induction prompts $p_v$ are removed, while all chain-of-thought reasoning $R_t$, actions $A_t^v$, and judge-verified observations (i.e., entries in $\mathcal{E}^v$) are preserved. The bridged context is forwarded to the Text Researcher, a deep research LLM equipped with web search tools, which generates the subsequent textual reasoning trajectory:

$$\begin{aligned}\tau_{\text{text}} = \{&D,\ q,\ R_1,\ A_1^v,\ O_1^v,\ \ldots,\ R_{T_v},\ A_{T_v}^v,\\ &O_{T_v}^v,\ R_{T_v+1},\ A_{T_v+1}^t,\ O_{T_v+1}^t,\\ &\ldots,\ R_{T_v+T_t},\ A_{T_v+T_t}^t\}\end{aligned} \tag{5}$$

where $T_t$ denotes the number of textual reasoning steps executed by the Text Researcher.

**Evidence Aggregation and Complete Trajectory Construction.** The Evidence Validator aggregates and verifies the complete collection of evidence produced across both reasoning phases. Taking as input all judge-verified visual observations in $\mathcal{E}^v$ and the textual reasoning trajectory $\tau_{\text{text}}$ produced by the Text Researcher, the Evidence Validator performs a cross-modal reasoning pass to resolve evidence conflicts, filter hallucinated content, and generate the final answer $a_{\text{out}}$. By grounding its reasoning in the graph-structured memory rather than a flattened context window, the Evidence Validator avoids the state-blindness inherent in linear memory paradigms, ensuring that the complete body of evidence across both modalities is faithfully incorporated into the final answer. The complete multimodal deep research trajectory is obtained by merging $\tau_{\text{vis}}$ and $\tau_{\text{text}}$:

$$\begin{aligned}\tau = \{&I,\ q,\ R_1,\ A_1^v,\ O_1^v,\ \ldots,\ R_{T_v},\ A_{T_v}^v,\\ &O_{T_v}^v,\ R_{T_v+1},\ A_{T_v+1}^t,\ O_{T_v+1}^t,\\ &\ldots,\ R_{T_v+T_t},\ A_{T_v+T_t}^t,\ a_{\text{out}}\}\end{aligned} \tag{6}$$

### 3.2 Multimodal Memory Graph

We construct a multimodal memory graph $\mathcal{G}_t = (\mathcal{V}_t,\ \mathcal{E}_t)$ that formalizes the multi-agent reasoning process as a dynamic DAG. This graph explicitly encodes action-observation dependencies across both the visual retrieval and textual research phases via a topological structure, fundamentally mitigating the state blindness inherent in linear memory paradigms and providing the structured foundation required by the TAPO algorithm. As illustrated in Figure 2(b), the memory graph evolves incrementally throughout reasoning, faithfully recording the exploration trajectory of each agent through node insertions and edge connections at each step.

**Node Schema and Modality Classification.** Each node $n_i \in \mathcal{V}_t$ encodes a discrete reasoning unit as a typed quadruple:

$$n_i \triangleq \big(\mu_i,\ \kappa_i,\ \sigma_i,\ \nu_i\big) \tag{7}$$

where $\mu_i \in \{n_{\text{root}},\ n_{\text{vis}},\ n_{\text{brd}},\ n_{\text{txt}},\ n_{\text{ans}}\}$ is the modality type tag; $\kappa_i$ is the search query or action descriptor driving the creation of this node; $\sigma_i$ is a concise textual summary of retrieved or generated content; and $\nu_i$ is the visual feature bank, populated only for $n_{\text{vis}}$-type nodes and set to empty for all other types. Five node types constitute the full reasoning pipeline. The root node ($n_{\text{root}}$) is initialized with user query $q$ and input image $I$, serving as the global anchor of the reasoning graph and permanently residing in $\mathcal{V}_t$. Visual nodes ($n_{\text{vis}}$) are spawned by the Visual Explorer at each retrieval step $t \in [1, T_v]$, storing the Hit/No Hit judgment and the corresponding evidence summary, with $\nu_i$ preserving multimodal content from matched web pages. The bridge node ($n_{\text{brd}}$) is a singleton node representing the modality bridging step, storing image description $D$ as the semantic interface between the two reasoning phases. Text nodes ($n_{\text{txt}}$) are spawned by the Text Researcher at each retrieval step $t \in [T_v + 1, T_v + T_t]$, storing web-retrieved textual evidence summaries. The answer node ($n_{\text{ans}}$) is spawned by the Evidence Validator, storing the final generated answer $a_{\text{out}}$, whose insertion marks the termination of the reasoning process. The edge set $\mathcal{E}_t$ connects temporally adjacent nodes , encoding the sequential dependency of the reasoning flow.

**Iterative Graph Expansion Protocol.** The memory graph $\mathcal{G}_t$ evolves through sequential node insertions jointly driven by the three main agents and the Hit/No Hit Judge. The expansion protocol consists of four phases as follows. During visual expansion ($t \in [1, T_v]$), the Visual Explorer executes retrieval action $A_t^v = \text{VSearch}(\mathcal{B}_t)$ and the Hit/No Hit Judge outputs binary signal $s_t$; a visual node $n_t = (n_{\text{vis}},\ \kappa_t,\ \sigma_t,\ \nu_t)$ is then inserted,

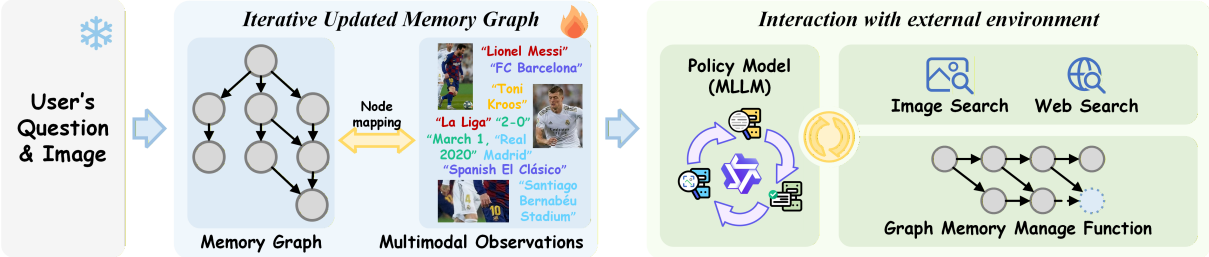


Figure 3: Inference pipeline of the GraMRAG framework.

where $\sigma_t$ encodes the evidence summary alongside the $s_t$ judgment, $\nu_t$ preserves visual content from matched web pages, and a sequential edge $(n_{t-1}, n_t)$ connects it to its predecessor. During modality bridging, upon termination of the visual phase, the bridge node $n_{\text{brd}} = (n_{\text{brd}}, \varnothing, D, \varnothing)$ is inserted and connected to the final visual node via the cross-modal bridge edge $(n_{T_v}, n_{\text{brd}})$, completing the semantic context transition from the visual to the textual domain. During textual expansion ($t \in [T_v + 1, T_v + T_t]$), the Text Researcher executes web retrieval actions; text nodes $n_t = (n_{\text{txt}}, \kappa_t, \sigma_t, \varnothing)$ are inserted with sequential edges, and the bridge node is connected to the first text node via edge $(n_{\text{brd}}, n_{T_v+1})$, fully closing the cross-modal transition path. During answer generation, once the Evidence Validator determines that the accumulated evidence is sufficient, it inserts the answer node $n_{\text{ans}} = (n_{\text{ans}}, \varnothing, a_{\text{out}}, \varnothing)$ and the reasoning process terminates. Upon completion of all four phases, the memory graph captures the complete multimodal reasoning trajectory as a structured DAG rather than compressing it into a one-dimensional sequence, thereby enabling each agent to efficiently query the full historical reasoning state and simultaneously providing the topological backbone required for fine-grained credit assignment in the reinforcement learning phase.

### 3.3 Topology-Aware Policy Optimization

As illustrated in Figure 4, we propose Topology-Aware Policy Optimization (TAPO) to address the systematic credit assignment problem in sparse reward reinforcement learning training, leveraging the topological structure of the memory graph to perform selective gradient masking that simultaneously rectifies both forms of bias across positive and negative trajectories with step-level precision.

**Graph-Structured Trajectory Decomposition.** We ground the reinforcement learning framework in the memory graph structure defined in Section 3.2. At each reasoning step $t$, the policy model $\pi_\theta$ receives the graph-serialized context:

$$C_t = \{s_{\text{sys}},\ q,\ H_t\} \tag{8}$$

where $s_{\text{sys}}$ is the system instruction and $H_t = \text{Serialize}(\mathcal{G}_{t-1})$ is the graph-linearized memory context. The policy model produces an reasoning segment that corresponds precisely to the construction of a single memory graph node $n_t$:

$$s_t = (C_t,\ \zeta_t,\ A_t,\ O_t) \longrightarrow n_t \tag{9}$$

where $\zeta_t$ denotes the chain-of-thought reasoning preceding action $A_t$, and $O_t$ is the environmental observation returned by the external tools. The terminal segment takes the form $s_{\text{ans}} = (C_{\text{ans}},\ \zeta_{\text{ans}},\ a_{\text{out}})$. A complete rollout trajectory decomposes into an ordered sequence of segments $\Psi = \{s_1,\ s_2,\ \ldots,\ s_{T_v+T_t},\ s_{\text{ans}}\}$. This node-based decomposition establishes a direct structural correspondence between policy gradient signals and graph topology: each gradient signal maps precisely to a specific graph node, providing the basis for the topology-driven credit assignment.

**Topology-Driven Credit Assignment.** For each rollout, we assign a binary outcome reward $r \in \{0, 1\}$ based on whether $a_{\text{out}}$ matches the ground-truth answer $a_{\text{true}}$. We then exploit the DAG topology of $\mathcal{G}_T$ to perform targeted gradient masking under two distinct scenarios as illustrated in Figure 4(b). Given a correct rollout ($r = 1$), we identify the critical reasoning path $\rho^*$ by traversing $\mathcal{G}_T$ backwards from $n_{\text{ans}}$ to $n_{\text{root}}$ along the directed edges. Nodes absent from $\rho^*$ constitute topological dead ends, namely exploratory steps that contribute no causal evidence to the correct answer. Propagating positive reward signals to these nodes would erroneously reinforce aimless exploration, so we suppress their gradient contributions. Given an incorrect rollout ($r = 0$), we identify the set of epistemically valuable nodes $\mathcal{N}_{\text{val}} \subseteq \mathcal{V}_T$. For $n_{\text{vis}}$-type

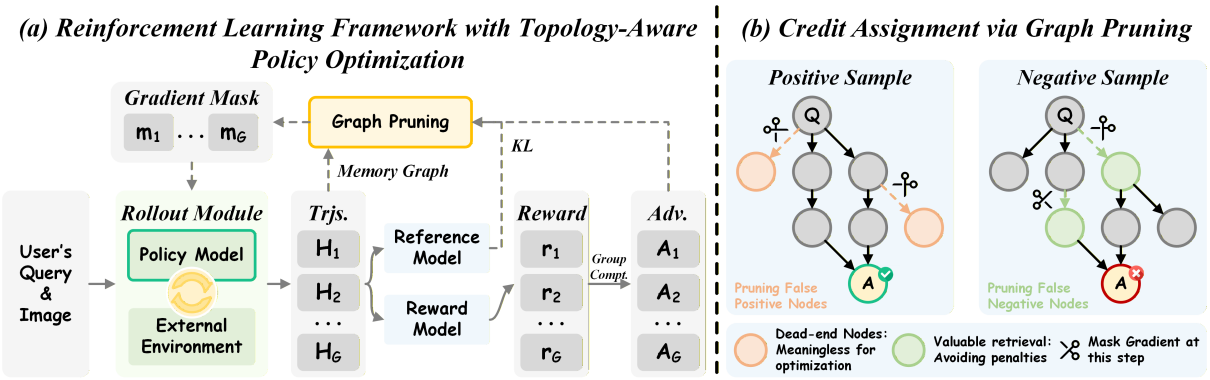


Figure 4: Overview of Topology-Aware Policy Optimization.

nodes, a node $n_t$ is designated as valuable when the Hit judgment $s_t$ stored within $\sigma_t$ records $s_t = 1$, confirming that the visual retrieval successfully located semantically relevant content. For $n_{\text{txt}}$-type nodes, a dedicated relevance evaluator further determines whether the textual evidence retrieved at step $t$ is directly relevant to the query. Excluding $\mathcal{N}_{\text{val}}$ from the negative policy gradient update prevents the model from being penalized from performing demonstrably productive retrievals on trajectories that ultimately fail due to shortcomings in other reasoning steps.

Integrating both scenarios, we define the gradient mask $\psi_t \in \{0, 1\}$ for segment $s_t$ as:

$$\psi_t = \mathbb{I}(r = 1) \cdot \mathbb{I}(n_t \notin \rho^*) + \mathbb{I}(r = 0) \cdot \mathbb{I}(n_t \in \mathcal{N}_{\text{val}}) \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function; $\psi_t = 1$ suppresses the gradient contribution of segment $s_t$ during the policy update, while $\psi_t = 0$ permits normal gradient update. The TAPO objective is defined as:

$$\mathcal{J}_{\text{TAPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{\sum_{g=1}^{G} n_g} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (1 - \psi_{g,i}) \cdot \mathcal{L}_{\text{clip}}\Big(\omega_{g,i}(\theta),\, \hat{A}_{g,i}\Big) - \beta \cdot \text{KL}[\pi_\theta \,\|\, \pi_{\text{ref}}]\Bigg] \tag{11}$$

where $\omega_{g,i}(\theta)$ is the probability ratio of segment $s_g^{(i)}$ under the current and old policies; $n_g$ denotes the number of segments in the $g$-th rollout. The $(1 - \psi_{g,i})$ factor selectively gates gradient contributions based on the topological analysis of the memory graph: in positive trajectories, topological dead-end steps are masked to prevent their gradients from reinforcing redundant exploration; in negative trajectories, valuable retrieval steps are masked to prevent their gradients from penalizing effective retrieval behavior. Through this design, TAPO achieves precise step-level credit assignment across the entire multi-step reasoning trajectory, significantly outperforming conventional sparse-reward methods that rely solely on terminal outcome signals.

# 4 Experiments

## 4.1 Datasets and Evaluation Metrics

We evaluate GraMRAG on a comprehensive set of benchmarks covering diverse task categories: general text benchmarks including HotpotQA (Yang et al., 2018) and SQuAD (Rajpurkar et al., 2016); image and visual document benchmarks including ViDoSeek (Wang et al., 2025a), SlideVQA (Tanaka et al., 2023) and MMLongBench (Ma et al., 2024); large-scale long-context video corpus benchmarks including LVBench (Wang et al., 2025b), WikiHowQA and SyntheticQA (Jeong et al., 2025), and XVBench (Wang et al., 2026). We employ a unified binary model-based metric (0 or 1) across all tasks, where a reward model performs semantic matching between the generated and reference answers to ensure consistent and fair evaluation.

## 4.2 Implementation Details

We conduct experiments using Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct (Bai et al., 2025) as the agent backbones. We use Qwen3-Max (Yang et al., 2025) as the reward model to evaluate the quality and relevance of generated responses. Training is divided into two stages. In the SFT stage, we employ LlamaFactory (Zheng et al., 2024) with LoRA (Hu et al., 2022) of rank 32 to fine-tune the language model while keeping the vision tower and multimodal projector frozen. Training spans 3 epochs with a batch size of 8 and gradient accumulation steps of 8, a learning rate of 1.0e-4, and a cosine learning rate scheduler. In the

| Method | General Text | | Image & Visual Document | | | Large-Scale Long-Context Video Corpus | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | HotpotQA | SQuAD | ViDoSeek | SlideVQA | MMLongBench | LVBench | WikiHowQA | SyntheticQA | XVBench | |
| *Qwen3-VL-4B-Instruct* | | | | | | | | | | |
| ReAct | 62.8 | 63.9 | 40.7 | 45.8 | 13.4 | 13.0 | 13.2 | 29.1 | 22.5 | 33.8 |
| Vanilla RAG | 62.4 | 63.2 | 44.9 | 43.4 | 14.9 | 12.8 | 12.1 | 33.7 | 27.1 | 34.9 |
| UniversalRAG | 51.7 | 65.3 | 40.3 | 14.6 | 5.2 | 14.9 | 3.8 | 22.9 | 6.9 | 25.1 |
| VideoRAG | 56.4 | 63.1 | 41.5 | 34.9 | 16.0 | 18.1 | 20.5 | 44.3 | 28.3 | 35.9 |
| MemAgent | 66.2 | 70.3 | 46.4 | 42.7 | 10.9 | 20.1 | 20.2 | 33.4 | 23.7 | 37.1 |
| Mem1 | 70.9 | 66.2 | 45.4 | 50.7 | 26.5 | 17.8 | 13.0 | 41.8 | 30.8 | 40.3 |
| VimRAG | 75.7 | 74.0 | 48.3 | 52.9 | 28.3 | 23.5 | 21.5 | 52.4 | 33.5 | 45.6 |
| **GraMRAG** | **78.4** | **77.1** | **51.2** | **55.6** | **30.8** | **26.0** | **24.1** | **57.3** | **36.0** | **48.5** |
| *Qwen3-VL-8B-Instruct* | | | | | | | | | | |
| ReAct | 70.7 | 64.7 | 41.3 | 50.4 | 14.7 | 15.8 | 21.7 | 34.2 | 25.1 | 37.6 |
| Vanilla RAG | 63.4 | 65.3 | 47.0 | 49.0 | 17.4 | 15.6 | 14.8 | 37.7 | 29.7 | 37.8 |
| UniversalRAG | 56.3 | 66.7 | 45.2 | 16.9 | 6.9 | 20.4 | 10.7 | 24.0 | 8.7 | 28.4 |
| VideoRAG | 62.5 | 62.3 | 43.1 | 36.0 | 19.3 | 23.8 | 27.1 | 48.6 | 32.1 | 39.4 |
| MemAgent | 72.3 | 74.1 | 48.1 | 44.8 | 13.5 | 23.0 | 23.4 | 36.8 | 28.3 | 40.5 |
| Mem1 | 71.5 | 68.6 | 44.2 | 57.0 | 34.0 | 22.5 | 20.8 | 44.0 | 31.4 | 43.8 |
| VimRAG | 78.6 | 76.6 | 53.3 | 63.1 | 34.7 | 23.2 | 29.0 | 54.5 | 37.7 | 50.1 |
| **GraMRAG** | **81.3** | **79.8** | **56.1** | **66.4** | **38.5** | **26.2** | **33.4** | **59.0** | **41.2** | **53.5** |

Table 1: Performance comparison across different benchmarks.

RL stage, we use rLLM (Tan et al., 2025) with the TAPO objective, an actor learning rate of 1.0e-6, a total of 2 training epochs, 8 agent groups, and a batch size of 32. All experiments are conducted on 8 NVIDIA H100 GPUs.

### 4.3 Comparison with State-of-the-Art Methods

We compare GraMRAG against representative baselines: ReAct (Yao et al., 2022) iteratively invokes external retrieval tools in a think-then-act paradigm; Vanilla RAG uses the original question as a query and performs direct MLLM inference over retrieved results; UniversalRAG (Yeo et al., 2025) formulates cross-modal retrieval as a unified routing problem; VideoRAG (Jeong et al., 2025) extracts significant visual information via key-frame selection; MemAgent (Yu et al., 2025) manages context by sequentially feeding search results to the model; Mem1 (Zhou et al., 2025) dynamically compresses agent memory through a cyclical retrieval-then-memorization process; and VimRAG (Wang et al., 2026) models reasoning process to manage large-scale multimodal contexts.

As presented in Table 1, linear history accumulation paradigms such as ReAct rapidly exhaust the context window under the heavy token overhead of visual data; task-specific methods including VideoRAG and UniversalRAG exhibit limited generalizability due to fixed inference pipelines; and summarization-based memory paradigms such as MemAgent and Mem1, lacking explicit modeling of agent reasoning states, progressively lose track of historical retrieval paths as interaction rounds accumulate; and VimRAG models the reasoning process but lacks long-horizon reasoning capability.

| Reasoning Horizon | | Memory Paradigm | | Policy Optimization | | Acc |
|---|---|---|---|---|---|---|
| Standard | Long | Iterative | Graph | GRPO | TAPO | |
| ✓ | | ✓ | | | | 43.8 |
| | ✓ | ✓ | | | | 44.2 |
| | ✓ | | ✓ | | | 47.5 |
| | ✓ | | ✓ | ✓ | | 49.3 |
| | ✓ | | ✓ | | ✓ | **53.5** |

Table 2: Ablation study across all benchmarks.

Our GraMRAG achieves consistently state-of-the-art performance across all benchmarks, achieving overall average scores of 48.5% and 53.5% under the Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct settings respectively, validating the effectiveness of formalizing the multi-step reasoning process as a dynamic DAG with topology-driven fine-grained credit assignment for multi-modal knowledge-intensive reasoning tasks.

### 4.4 Ablation Study

To validate the contribution of each component in GraMRAG, we conduct ablation studies across all benchmarks. As shown in Table 2, the three components exhibit clear progressive complementarity: long-horizon reasoning enables sustained multi-round retrieval across extended cross-modal reasoning trajectories; the multimodal memory graph demonstrates that formalizing reasoning as a dynamic DAG effectively mitigates state blindness and suppresses redundant retrieval; and replacing the standard sparse-reward method Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with TAPO delivers a further improvement, directly confirming the critical role of topology-guided credit assignment in fine-grained policy optimization. The complete model achieves an average score of 53.5%, an improvement of over 9 points

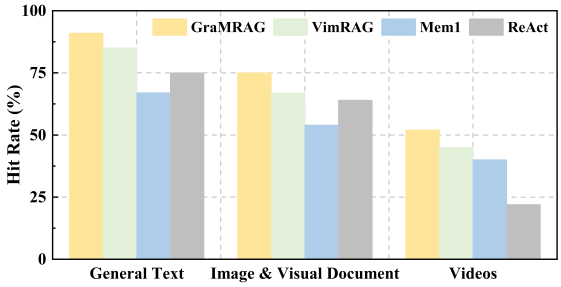


Figure 5: Analysis of retrieval performance.

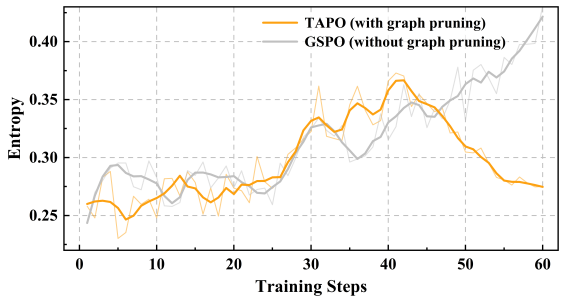


Figure 6: Analysis of training efficiency.

above the baseline, comprehensively validating the necessity and complementarity of each component.

## 5 Analysis and Discussion

**Retrieval quality depends on reasoning depth.** High-quality answer generation in complex multimodal tasks depends on sustaining precise multi-round retrieval across a long reasoning horizon. Traditional memory paradigms cause agents to progressively lose track of historical retrieval paths and accumulate redundant searches that severely constrain effective reasoning depth. GraMRAG addresses this by transferring the iterative planning capability into the multimodal domain via the vision-text bridged reasoning paradigm, supporting trajectories spanning tens of steps; the multimodal memory graph further encodes action-observation dependencies explicitly, directing each retrieval attempt precisely toward unexplored evidence space. As shown in Figure 5, this design yields substantially higher hit rate across all three task categories, directly validating the decisive role of long-horizon reasoning depth in determining retrieval quality.

**Graph memory mitigates state blindness.** The dynamic DAG representation of the multimodal memory graph explicitly encodes action-observation dependencies as directed edges, enabling each agent to efficiently access the complete historical reasoning trajectory and effectively mitigating state blindness. As shown in Table 2, replacing iterative summarization-based memory with the multimodal memory graph yields a 3.3-point performance improvement, validating that the multimodal memory graph effectively mitigates state blindness and suppresses redundant retrieval, confirming the necessity of structured memory in long-horizon multimodal reasoning.

**Graph pruning accelerates training convergence.** As demonstrated in Figure 6, TAPO achieves faster convergence throughout training compared to the baseline Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025), indicating that topology-guided graph pruning drives the model toward faster convergence by supplying cleaner credit assignment signals. This yields two critical insights into agentic reinforcement learning: first, optimization stability fundamentally depends on ensuring correct positive gradient signals while eliminating ambiguous updates from negative trajectories; second, the quality of rollout samples, specifically their preference alignment, is more decisive for performance than simply scaling up the training set. By simultaneously masking dead-end nodes in positive trajectories and preserving valuable nodes in negative ones, TAPO achieves fine-grained step-level credit assignment rather than coarse terminal-reward-level optimization.

## 6 Conclusion

In this paper, we introduce GraMRAG, a novel graph memory-guided multi-agent RAG framework for complex multimodal long-horizon reasoning tasks. Our approach integrates three key innovations: a vision-text bridged reasoning paradigm that unifies multi-scale entity cropping with a ReAct-style visual toolchain for long-horizon cross-modal reasoning; multimodal memory graph construction that formalizes agent reasoning as a dynamic directed acyclic graph (DAG) to fundamentally mitigate state blindness and suppress redundant retrieval; and Topology-Aware Policy Optimization (TAPO) algorithm that leverages graph topology for fine-grained credit assignment across multi-step reasoning trajectories. Extensive experiments on diverse multimodal benchmarks demonstrate that our method consistently outperforms existing baselines and achieves state-of-the-art performance. Future work will investigate the scalability of the proposed framework to larger architectures and broader multimodal scenarios.

## 7 Limitations

Although GraMRAG achieves significant performance improvements on complex multimodal multi-step reasoning tasks, the multi-agent collaboration framework introduces some inference overhead compared to traditional RAG approaches. Notably, such a trade-off is generally acceptable in deep research scenarios where users prioritize reasoning accuracy over response speed. Nevertheless, it may hinder direct deployment in resource-constrained environments. Therefore, optimizing the interaction strategy among agents and designing more efficient lightweight graph memory mechanisms constitute key directions for future work.

## References


Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and 1 others. 2025. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Jiangwen Dong, Zehui Lin, Wanyu Lin, and Mingjin Zhang. 2026. S-dag: A subject-based directed acyclic graph for multi-agent heterogeneous reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29394–29402.

Xuanbo Fan, Tianqi Zhao, Yi Cheng, Chi Xiu, Jiaxin Guo, Boci Peng, Bingjing Xu, Jessica Zhang, Feng Sun, and Yan Zhang. 2026. Segmem-rag: Adaptive memory for retrieval-augmented generation in open-ended knowledge environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30647–30655.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Ziyi Guan, Jason Chun Lok Li, Zhijian Hou, Pingping Zhang, Donglai Xu, Yuzhi Zhao, Mengyang Wu, Jinpeng Chen, Thanh-Toan Nguyen, Pengfei Xian, and 1 others. 2025. Kg-rag: Enhancing gui agent decision-making via knowledge graph-driven retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5396–5405.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Hao Hu, Yifan Feng, Ruoxue Li, Rundong Xue, Xingliang Hou, Zhiqiang Tian, Yue Gao, and Shaoyi Du. 2026. Cog-rag: Cognitive-inspired dual-hypergraph with theme alignment retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31032–31040.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, and 1 others. 2026. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21278–21298.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, and 1 others. 2025. Mmsearch: Unveiling the potential of large models as multimodal search engines. In *The Thirteenth International Conference on Learning Representations*.

Yuanhao Li, Mingshan Liu, Hongbo Wang, Yiding Zhang, Yifei Ma, and Wei Tan. 2026. Draft-rl: Multi-agent chain-of-draft reasoning for reinforcement learning-enhanced llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29530–29537.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. 2025. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*.

Yiran Ni and Deshi Ye. 2026. Multi-agent corridor reasoning for multi-agent path finding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29617–29624.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shikhhar Siingh, Abhinav Rawat, Chitta Baral, and Vivek Gupta. 2025. Getreason: Enhancing image context extraction through hierarchical multi-agent reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29779–29800.

Sanchit Sinha, Guangzhi Xiong, Zhenghao He, and Aidong Zhang. 2026. Concept-rulenet: Grounded multi-agent neurosymbolic reasoning in vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29678–29686.

Sijun Tan, Michael Luo, Colin Cai, Tarun Venkat, Kyle Montgomery, Aaron Hao, Tianhao Wu, Arnav Balyan, Manan Roongta, Chenguang Wang, and 1 others. 2025. rllm: A framework for post-training language agents.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025a. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9124–9145.

Qiuchen Wang, Shihang Wang, Yu Zeng, Qiang Zhang, Fanrui Zhang, Zhuoning Guo, Bosi Zhang, Wenxuan Huang, Lin Chen, Zehui Chen, and 1 others. 2026. Vimrag: Navigating massive visual context in retrieval-augmented generation via multimodal memory graph. *arXiv preprint arXiv:2602.12735*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025b. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025a. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, and 1 others. 2025b. Resum: Unlocking long-horizon search intelligence via context summarization. *arXiv preprint arXiv:2509.13313*.

Zhao Xinjie, Fan Gao, Xingyu Song, Yingjian Chen, Rui Yang, Yanran Fu, Yuyang Wang, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Reagent: Reversible multi-agent reasoning for knowledge-enhanced multi-hop qa. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4067–4089.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025a. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025b. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, and 1 others. 2025. Agentfold: Long-horizon web agents with proactive context management. *arXiv preprint arXiv:2510.24699*.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. 2025. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and 1 others. 2025. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *arXiv preprint arXiv:2507.02259*.

Taolin Zhang, Dongyang Li, Qizhou Chen, Chengyu Wang, and Xiaofeng He. 2025a. Belle: A bi-level multi-agent reasoning framework for multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4184–4202.

Wenchuan Zhang, Jingru Guo, Hengzhe Zhang, Penghao Zhang, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. 2026. Patho-agenticrag: towards multimodal agentic retrieval-augmented generation for pathology vlms via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29921–29929.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025b. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.

Yuxiang Zhou, Jichang Li, Yanhao Zhang, Haonan Lu, and Guanbin Li. 2026. Mobile-agent-rag: Driving smart multi-agent coordination with contextual knowledge empowerment for long-horizon mobile automation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29939–29947.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*.