MG^2-RAG: Multi-Granularity Graph for Multimodal Retrieval-Augmented Generation

Sijun Dai[1,†], Qiang Huang[1,†], Xiaoxing You[2], and Jun Yu[1,*]

[1] School of Intelligence Science and Engineering,
Harbin Institute of Technology (Shenzhen)
daisijun@stu.hit.edu.cn, {huangqiang, yujun}@hit.edu.cn
[2] School of Computer Science, Hangzhou Dianzi University
youxiaoxing@hdu.edu.cn
[†]Equal contribution [*]Corresponding author

Abstract. Retrieval-Augmented Generation (RAG) mitigates hallucinations in Multimodal Large Language Models (MLLMs), yet existing systems struggle with complex cross-modal reasoning. Flat vector retrieval often ignores structural dependencies, while current graph-based methods rely on costly "translation-to-text" pipelines that discard fine-grained visual information. To address these limitations, we propose **MG^2-RAG**, a lightweight **M**ulti-**G**ranularity **G**raph **RAG** framework that jointly improves graph construction, modality fusion, and cross-modal retrieval. MG^2-RAG constructs a hierarchical multimodal knowledge graph by combining lightweight textual parsing with entity-driven visual grounding, enabling textual entities and visual regions to be fused into unified multimodal nodes that preserve atomic evidence. Building on this representation, we introduce a multi-granularity graph retrieval mechanism that aggregates dense similarities and propagates relevance across the graph to support structured multi-hop reasoning. Extensive experiments across four representative multimodal tasks (i.e., retrieval, knowledge-based VQA, reasoning, and classification) demonstrate that MG^2-RAG consistently achieves state-of-the-art performance while reducing graph construction overhead with an average 43.3× speedup and 23.9× cost reduction compared with advanced graph-based frameworks. The source code is publicly available at https://github.com/Daboolu/MG2-RAG.

Keywords: Multimodal Knowledge Graph · Retrieval-Augmented Generation · Multimodal Large Language Models

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable success in tasks requiring complex cross-modal understanding and reasoning, leading to their rapid adoption across numerous applications [20, 27, 62, 64]. Despite these advances, deploying MLLMs in knowledge-intensive scenarios still raises important reliability concerns. In particular, MLLMs may produce multimodal hallucinations and often lack access to private or domain-specific knowledge, since

their parameters are primarily trained on large-scale public corpora [34, 72, 86]. To address these limitations, Multimodal Retrieval-Augmented Generation (MM-RAG) augments MLLMs with external knowledge bases by retrieving relevant multimodal evidence to ground the generation process [5, 33, 77, 81, 84]. By incorporating factual context at inference time, MM-RAG significantly improves both the reliability and domain adaptability of MLLMs.

Existing MM-RAG methods generally fall into two paradigms: vector-based and graph-based approaches. Vector-based MM-RAG encodes multimodal inputs into a shared embedding space [61,65,68] and retrieves relevant evidence via similarity search [12, 13]. While effective in many scenarios, this paradigm typically retrieves isolated multimodal elements and largely ignores the structural relationships among different pieces of evidence [9, 32, 48, 74]. Consequently, two fundamental limitations arise: First, the semantic gap between modalities can exacerbate modality fragmentation, weakening cross-modal alignment. Second, similarity-based retrieval cannot model logical dependencies among retrieved evidence, making complex multi-hop reasoning difficult.

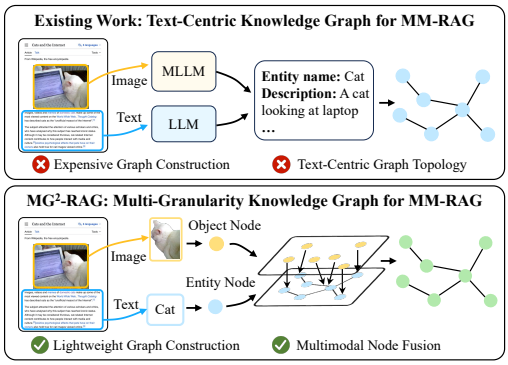

Fig. 1: Comparison between existing graph-based MM-RAG frameworks and MG^2-RAG. Existing methods rely on costly text-centric graph construction that discards fine-grained visual information, whereas MG^2-RAG efficiently fuses textual entities and visual objects into unified multimodal nodes while preserving atomic evidence.

To overcome these limitations, recent work has explored graph-based MM-RAG, which organizes multimodal knowledge into structured graphs with explicit relational dependencies. By representing entities and their relationships as graph nodes and edges, these approaches enable retrieval through structured traversal rather than flat similarity matching [51, 69, 76, 79]. Such structured representations allow models to reason over interconnected multimodal facts and improve reliability in knowledge-intensive tasks [23, 25, 87]. As a result, graph-based retrieval has emerged as a promising direction for bridging multimodal alignment and structured reasoning. Despite its promise, existing graph-based MM-RAG systems still face several practical challenges.

- **Expensive Graph Construction:** Many existing methods rely heavily on MLLMs to extract relational triplets from multimodal data, resulting in significant overhead when processing large-scale knowledge bases [51, 69, 79].
- **Text-Centric Graph Topology:** Visual information is often converted into textual descriptions before constructing the graph [23, 69, 87]. This transformation discards fine-grained visual structures and overlooks independent visual entities that may contain important semantic information.

– **Difficult Cross-Modal Retrieval:** Since these graphs lack unified multimodal concepts, they struggle to effectively process queries containing both text and images, which limits semantic propagation across modalities and weakens multi-hop reasoning capabilities [51, 74].

To address these challenges, we propose **MG^2-RAG**, a lightweight **M**ulti-**G**ranularity **G**raph **RAG** framework for efficient and reliable multimodal retrieval-augmented generation. MG^2-RAG constructs hierarchical multimodal knowledge graphs by combining lightweight textual parsing with entity-driven visual grounding, avoiding expensive MLLM-driven graph construction. It further fuses textual entities and visual objects into unified multimodal nodes while preserving fine-grained atomic evidence from both modalities.

Building upon this representation, we introduce a multi-granularity graph retrieval mechanism that aggregates dense similarities onto multimodal nodes and propagates relevance through the graph topology, enabling unified multimodal retrieval and structured multi-hop reasoning. Extensive experiments on four representative multimodal tasks, including retrieval, knowledge-based VQA, reasoning, and classification, demonstrate that MG^2-RAG consistently achieves state-of-the-art performance while substantially reducing graph construction overhead. Our main contributions are summarized as follows:

– **Lightweight Multimodal Knowledge Graph Construction.** We propose a lightweight pipeline that bypasses the expensive MLLM-driven triplet extraction process, significantly reducing the time and cost of multimodal knowledge graph construction.
– **Modality-Preserving Multimodal Node Fusion.** We introduce a modality-preserving fusion strategy that integrates textual entities and visual objects into unified multimodal nodes while preserving fine-grained atomic evidence from both modalities.
– **Multi-Granularity Graph Retrieval.** We design a retrieval mechanism that aggregates dense similarities onto multimodal nodes and propagates relevance through the graph topology, enabling unified multimodal retrieval and multi-hop reasoning for MLLM generation.

2 Related Work

RAG for LLMs. RAG has evolved from static *retrieve-then-generate* pipelines to *dynamic, agentic reasoning* systems. Early approaches mitigate hallucinations by augmenting language models with external knowledge, progressing from sparse Wikipedia retrieval [7,40] to dense retrieval [26,38,43]. Subsequent work introduces adaptive retrieval strategies that enable models to determine when and what to retrieve, often optimized through reinforcement learning and environmental feedback [2, 11, 19, 31]. Recently, RAG has become tightly coupled with multi-hop reasoning through hierarchical query decomposition [41, 83], iterative retrieval-reasoning loops [22,70], and agentic workflows [42,47,73]. These advances improve performance on complex, knowledge-intensive tasks. Nonetheless, most existing methods still rely on *flat vector retrieval*, limiting their ability to model structured semantic dependencies and motivating graph-based retrieval paradigms.

Graph RAG. To overcome the structural limitations of flat retrieval, recent work has incorporated graph-based representations and decoupled subgraph retrieval into RAG [16, 29, 56, 66, 77, 80]. Graph RAG frameworks leverage entity-level knowledge graphs and community structures to better capture global dependencies and multi-hop relationships [16,17,76]. Given the high computational cost of graph construction, subsequent research has focused on scalable indexing and retrieval strategies, including dual-level retrieval [23], bipartite or relation-free graph designs [35, 87], lightweight parallel triple-scoring [46], and hierarchical tree-based structures [71, 85]. Beyond efficiency, integrating graphs with LLMs has proven crucial for faithful multi-hop reasoning and long-term memory modeling. Iterative graph exploration [36, 58, 67, 78], planning over structured representations [8, 55], memory-inspired architectures [24, 25], and agentic frameworks [15] demonstrate that explicit relational modeling significantly enhances reasoning depth and interpretability. Nevertheless, these methods are predominantly *text-centric* and are not directly designed for multimodal settings.

Multimodal RAG. Vision-language pre-training models [45, 65] have significantly advanced multimodal understanding, yet they remain constrained by parametric knowledge. Multimodal RAG (MM-RAG) addresses this limitation by introducing external memory for grounding generation in retrieved visual–textual evidence [9, 33, 75]. Recent efforts improve retrieval granularity [18, 57] and incorporate adaptive or self-reflective strategies [5, 12, 82] to filter noise and control retrieval necessity. However, most systems rely on flat vector similarity, which struggles with structured multi-hop reasoning across modalities [37, 52].

To address this, Multimodal Knowledge Graphs (MMKGs) [39, 51, 79] have emerged as a promising direction for structured multimodal retrieval. For example, MMGraphRAG [69] leverages scene graphs and clustering to connect visual and textual information. Yet existing methods typically rely on MLLMs to translate visual content into textual graph nodes, yielding text-centric graph representations that discard fine-grained visual structures. Consequently, preserving atomic visual regions together with textual entities in a **natively unified multimodal graph** remains largely underexplored.

3 Methodology

We propose **MG^2-RAG**, a lightweight framework that transforms an unstructured multimodal knowledge base into a multi-granularity MMKG for retrieval-augmented generation. As shown in Figure 2, the framework consists of two modules: (1) Hierarchical Multimodal Knowledge Graph Construction and (2) Multi-Granularity Graph Retrieval.

3.1 Hierarchical Multimodal Knowledge Graph Construction

Motivation. Existing MMKG construction strategies suffer from two key issues: (1) expensive LLM-based relation extraction and (2) modality fragmentation caused by converting images into textual descriptions. To address these issues,

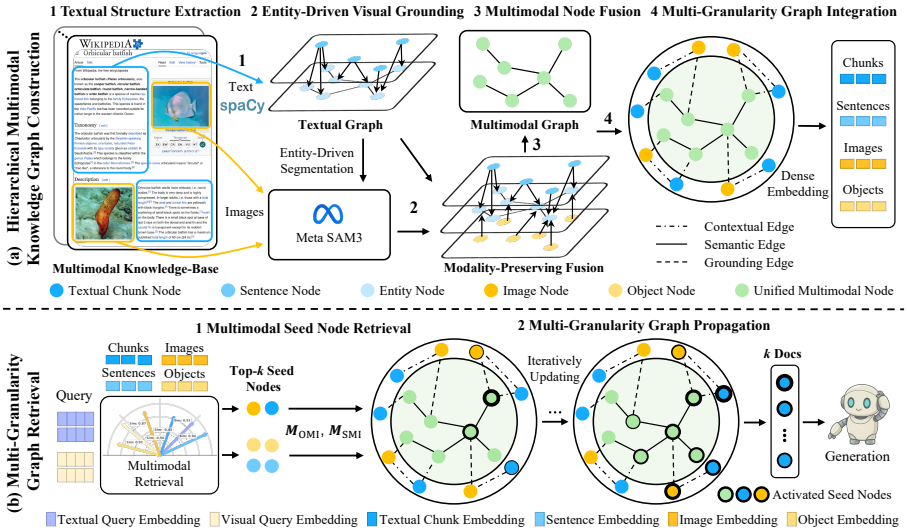

Fig. 2: Overview of MG^2-RAG. MG^2-RAG contains two modules: **(a) Hierarchical Multimodal Knowledge Graph Construction** that transforms a multimodal knowledge base into a hierarchical fine-grained multimodal knowledge graph (MMKG) via textual parsing and entity-driven visual grounding; **(b) Multi-Granularity Graph Retrieval** that employs a multi-granularity retrieval mechanism, driven by graph propagation, to retrieve the top-K most relevant document chunks for MLLM generation.

we construct a hierarchical MMKG that directly aligns textual entities with grounded visual objects, forming atomic unified multimodal nodes. We define the multimodal knowledge base as $\mathcal{D} = \{(c_k, \mathcal{I}_k)\}_{k=1}^{N}$, where each unit consists of a textual document chunk c_k and its associated image set $\mathcal{I}_k = \{I_{k,1}, \cdots, I_{k,m_k}\}$. Rather than modeling text–image correspondence as loose cross-modal links, we explicitly extract and fuse semantic information to construct a compact hierarchical MMKG. Each unified multimodal node preserves fine-grained textual and visual components internally, enabling nuanced downstream reasoning.

Textual Structure Extraction: NER and Dependency Parsing. We first extract structural signals from text using the transformer-based spaCy model `en_core_web_trf` [30]. Each chunk c_k is parsed into a sentence set $\mathcal{S}_k = \{s_{k,1}, \cdots, s_{k,n_k}\}$, from which we extract a local entity set $\mathcal{E}_k$. Aggregating across all chunks yields the global sentence and entity sets:

$$\mathcal{S} = \bigcup_{k=1}^{N} \mathcal{S}_k, \quad \mathcal{E} = \bigcup_{k=1}^{N} \mathcal{E}_k.$$

To capture structural relations between entities, we leverage token-level dependency parsing. Instead of relying on expensive LLM-based OpenIE extraction [15, 24], we adopt a *lightweight* rule-based relation extraction strategy grounded in dominant grammatical patterns (e.g., predicate-centered and nominal-modifier structures). This significantly reduces construction cost while preserving high-precision relational structure. Details are provided in Appendix A.1.

Entity-Driven Visual Grounding. To prevent the "translation-to-text" bottleneck, we directly ground textual entities in visual space. Given the local entity set $\mathcal{E}_k$ and image set $\mathcal{I}_k$ for each chunk c_k, we perform *entity-driven segmentation* rather than global image parsing. Specifically, we batch entity names as semantic prompts into an open-vocabulary segmentation model Φ (SAM3 [6]) to localize entity-relevant regions. For entity $e \in \mathcal{E}_k$, its grounded region set is:

$$\mathcal{V}_o^e = \{o \mid (o, \sigma) \in \Phi(e, I),\ I \in \mathcal{I}_k,\ \sigma > \tau\}, \tag{1}$$

where o is the segmented object, σ denotes grounding confidence for entity e on image I, and τ is a threshold. This targeted grounding ensures that visual objects are directly and semantically aligned with textual entities.

Modality-Preserving Multimodal Node Fusion. For entities with valid detections ($\mathcal{V}_o^e \neq \emptyset$), we construct a modality-preserving multimodal node $v_m = (e, \mathcal{V}_o^e) \in \mathcal{V}_M$, where $\mathcal{V}_M$ denotes the set of multimodal nodes. To explicitly model internal composition, we define:

- **Object-Multimodal Incidence Matrix:** $\boldsymbol{M}_{\text{OMI}} \in \{0,1\}^{|\mathcal{V}_O| \times |\mathcal{V}_M|}$, where $\mathcal{V}_O = \bigcup_{e \in \mathcal{E}} \mathcal{V}_o^e$. Entry $\boldsymbol{M}_{\text{OMI}}[i,j] = 1$ indicates that object $o_i \in \mathcal{V}_O$ belongs to multimodal node $v_{m_j} \in \mathcal{V}_M$.
- **Sentence-Multimodal Incidence Matrix:** $\boldsymbol{M}_{\text{SMI}} \in \{0,1\}^{|\mathcal{S}| \times |\mathcal{V}_M|}$, where $\boldsymbol{M}_{\text{SMI}}[i,j] = 1$ indicates entity e_j that is associated with the unified multimodal node v_{m_j} appears in sentence $s_i \in \mathcal{S}$.

These matrices enable structured aggregation during retrieval.

Multi-Granularity Graph Integration. Building upon the matrices of document chucks $\mathcal{V}_C$, images $\mathcal{V}_I$, and multimodal nodes $\mathcal{V}_M$, we construct multi-granularity multimodal graph:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}_{\mathcal{G}}), \quad \mathcal{V} = \mathcal{V}_C \cup \mathcal{V}_I \cup \mathcal{V}_M.$$

For each multimodal unit $(c_k, \mathcal{I}_k)$, nodes operate at three levels: chunk nodes $v_c^k \in \mathcal{V}_C$, image nodes $v_I^{k,j} \in \mathcal{V}_I$ for each associated image $I_{k,j} \in \mathcal{I}_k$, and unified multimodal nodes $v_m \in \mathcal{V}_M$. To represent rich structural and semantic interactions, the edge set $\mathcal{E}_{\mathcal{G}}$ is constructed across three complementary dimensions:

(1) **Contextual Edges:** To preserve hierarchical provenance, each chunk node v_c^k is linked to its associated image nodes $v_I^{k,j}$ and multimodal nodes v_m derived from $e \in \mathcal{E}_k$. These edges maintain document-level structural coherence.

(2) **Semantic Edges:** To encode linguistic structure, we connect multimodal nodes (v_{m_a}, v_{m_b}) when their underlying textual entities exhibit direct syntactic dependencies. This enables structured reasoning over explicit relational cues extracted from text.

(3) **Grounding Edges:** To model explicit cross-modal alignment, each visual object $o \in \mathcal{V}_o^e$ encapsulated within a multimodal node v_m is linked to its source image node $v_I^{k,j}$. The edge weight is assigned as the grounding confidence σ, reflecting the reliability of the localized visual evidence.

These three edge types integrate contextual hierarchy, semantic structure, and visual grounding into a unified heterogeneous topology, laying the foundation for structured multi-granularity retrieval in Section 3.2

In parallel, all textual and visual elements are embedded into a shared space via a unified encoder $\Psi(\cdot)$ (i.e., `EVA-CLIP-8B` [68]),[3] which maps any elemental input x to a d-dimensional vector $\boldsymbol{z}_x = \Psi(x) \in \mathbb{R}^d$. We construct four embedding matrices corresponding to distinct levels of granularity: (1) **Sentence matrix** $\boldsymbol{Z}_S \in \mathbb{R}^{|\mathcal{S}| \times d}$ encoded from global sentences $\mathcal{S}$; (2) **Chunk matrix** $\boldsymbol{Z}_C \in \mathbb{R}^{|\mathcal{V}_C| \times d}$ encoded from document chunks $\mathcal{V}_C$; (3) **Image matrix** $\boldsymbol{Z}_I \in \mathbb{R}^{|\mathcal{V}_I| \times d}$ encoded from source images $\mathcal{V}_I$; and (4) **Object matrix** $\boldsymbol{Z}_O \in \mathbb{R}^{|\mathcal{V}_O| \times d}$ encoded from segmented objects $\mathcal{V}_O$. Together, the graph topology and dense embeddings enable multi-granularity retrieval.

3.2 Multi-Granularity Graph Retrieval

Motivation. Flat vector retrieval treats modalities independently and ignores structural dependencies within multimodal knowledge bases. To enable structured cross-modal reasoning, we adopt a two-stage strategy: (1) activating fine-grained multimodal evidence as seed nodes, and (2) propagating their relevance across the graph topology to capture multi-hop interactions.

Multimodal Seed Node Retrieval. Leveraging the multimodal index constructed in Section 3.1, a textual query q_t or visual query q_v retrieves candidate evidence from all semantic levels. For each modality $m \in \{t, v\}$, dense similarities with the embedding matrices $\boldsymbol{Z}_S$, $\boldsymbol{Z}_C$, $\boldsymbol{Z}_I$, and $\boldsymbol{Z}_O$ produce four relevance vectors: $\boldsymbol{s}_S^{(m)} \in \mathbb{R}^{|\mathcal{S}|}$, $\boldsymbol{s}_C^{(m)} \in \mathbb{R}^{|\mathcal{V}_C|}$, $\boldsymbol{s}_I^{(m)} \in \mathbb{R}^{|\mathcal{V}_I|}$, and $\boldsymbol{s}_O^{(m)} \in \mathbb{R}^{|\mathcal{V}_O|}$.

Because sentences $\mathcal{S}$ and visual objects $\mathcal{V}_O$ act as intermediate pivots rather than primary graph nodes, their scores are aggregated onto unified multimodal nodes $\mathcal{V}_M$ using the incidence matrices $\boldsymbol{M}_{\text{SMI}}$ and $\boldsymbol{M}_{\text{OMI}}$:

$$\boldsymbol{s}_M^{(m)} = \boldsymbol{D}_S^{-1} \boldsymbol{M}_{\text{SMI}}^\top \boldsymbol{s}_S^{(m)} + \boldsymbol{D}_O^{-1} \boldsymbol{M}_{\text{OMI}}^\top \boldsymbol{s}_O^{(m)}, \tag{2}$$

where $\boldsymbol{D}_S, \boldsymbol{D}_O \in \mathbb{R}^{|\mathcal{V}_M| \times |\mathcal{V}_M|}$ are diagonal degree matrices performing mean pooling based on the number of mapped unified multimodal nodes.

To balance evidence from different semantic levels, we introduce scaling factors ω_C and ω_I for chunk and image nodes. Given the node partition $\mathcal{V} = \mathcal{V}_C \cup \mathcal{V}_I \cup \mathcal{V}_M$, the modality-specific activation vector is $\boldsymbol{u}^{(m)} = [\,\omega_C \boldsymbol{s}_C^{(m)};\ \omega_I \boldsymbol{s}_I^{(m)};\ \boldsymbol{s}_M^{(m)}\,]$. Finally, modality contributions are fused using weights λ_t and λ_v:

$$\boldsymbol{r}_0 = \lambda_t \boldsymbol{u}^{(t)} + \lambda_v \boldsymbol{u}^{(v)}. \tag{3}$$

To reduce noise before graph propagation, we retain only the top-k activated nodes and normalize the resulting distribution, as detailed in Appendix A.3.

Multi-Granularity Graph Propagation. Starting from the seed distribution $\boldsymbol{r}_0$, we propagate relevance across the heterogeneous graph using Personalized PageRank [28]. Let $\boldsymbol{W}$ denote the column-normalized transition matrix derived from $\mathcal{E}_\mathcal{G}$. The diffusion process iteratively updates node activations as:

$$\boldsymbol{r}_{\ell+1} = \alpha \boldsymbol{W} \boldsymbol{r}_\ell + (1 - \alpha) \boldsymbol{r}_0, \tag{4}$$

[3] Without specialization, all embeddings are explicitly ℓ_2-normalized, and their semantic proximity is computed directly via cosine similarity.

until $\Delta \boldsymbol{r} = \boldsymbol{r}_{\ell+1} - \boldsymbol{r}_\ell \leq \epsilon$, where $\alpha \in (0, 1)$ controls the propagation strength and $(1 - \alpha)$ ensures periodic restarts from the seed distribution $\boldsymbol{r}_0$. After convergence, the relevance scores of chunk nodes are extracted to select the top-k document chunks. These chunks are combined with the multimodal query (q_t, q_v) to form the final prompt for the MLLM, enabling grounded response generation.

4 Experiment

4.1 Experimental Setup

Baselines. We compare **MG^2-RAG** against three state-of-the-art graph-based MM-RAG frameworks: **VaLiK** [51], **mKG-RAG** [79], and **MMGraphRAG** [69]. For fair comparison, we report results using the best-performing configurations from the original papers without dataset-specific fine-tuning. Additional comparisons with other relevant approaches are summarized in Tables 2–5.

Multimodal Tasks, Datasets, and Evaluation Metrics. We evaluate MG^2-RAG on four representative multimodal tasks, covering retrieval, knowledge-based Visual Question Answering (VQA), reasoning, and classification.

- **Multimodal Retrieval:** We evaluate retrieval performance on two Wikipedia-linked datasets: **E-VQA** [59] and **InfoSeek** [10]. E-VQA contains diverse fine-grained entities and includes **single-hop** and **two-hop** multi-page queries, making retrieval challenging. InfoSeek provides two evaluation splits: Unseen Entity (**Unseen-E**) and Unseen Question (**Unseen-Q**), where the former requires retrieving previously unseen entities. For both datasets, we construct MMKGs from the associated corpora and retrieve supporting evidence for each query. Performance is evaluated using Recall@K (**R@K**).
- **Knowledge-based VQA:** Using the retrieved evidence, the model generates answers for both textual and visual queries. In **E-VQA**, questions may involve multiple associated images and require reasoning across one or two knowledge hops. For **InfoSeek**, evaluation on both Unseen-E and Unseen-Q assesses generalization to novel entities and question formulations. Performance is measured using BERT Matching Score (**BEM**) [4] for E-VQA, and **standard accuracy** [21] or **relaxed accuracy** [60] for InfoSeek.
- **Multimodal Reasoning:** We further evaluate our method on **ScienceQA** [53], which contains around 21k multimodal science questions spanning physics, chemistry, and biology. Following VaLiK [51], the training set is used as the knowledge base to construct the MMKG. The task requires answering complex scientific questions through multimodal reasoning. Performance is measured using **classification accuracy**, reported across different question types, contextual modalities, and educational stages.
- **Multimodal Classification:** To assess real-world applicability, we evaluate MG^2-RAG on **CrisisMMD** [1], a disaster-response dataset containing approximately 35k noisy social-media image-text pairs. Following VaLiK [51], the training set is used to construct the MMKG, capturing modality correlations and noise patterns typical of user-generated content. We evaluate three

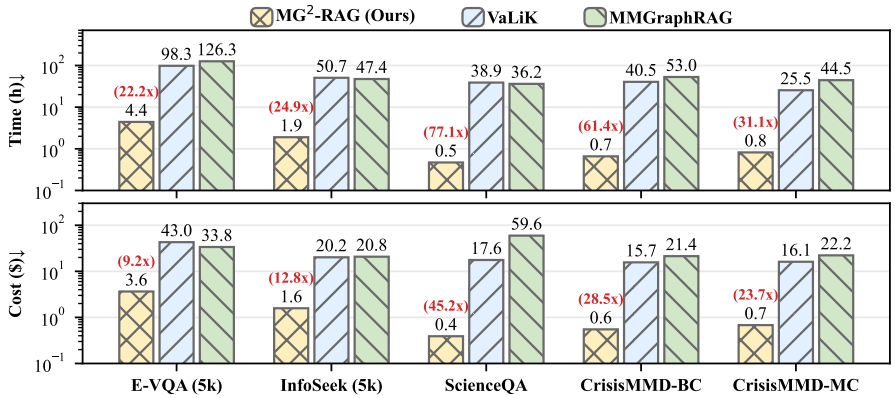

Fig. 3: Graph construction efficiency. The top and bottom plots show construction time (hours) and cost ($), respectively. Red annotations above the MG^2-RAG bars indicate the efficiency gain (speedup or cost reduction) relative to the strongest baseline.

classification tasks: (1) **BC**: binary classification for information relevance filtering, (2) **MC**: multi-class classification of fine-grained humanitarian categories, and (3) **MC-m**: a simplified multi-class setting with merged categories. Performance is measured using **standard classification accuracy**.

Further details on dataset preprocessing are provided in Appendix A.2.

Implementation Details. For **Hierarchical MMKG Construction**, textual analysis is performed using spaCy (`en_core_web_trf`) [30], and entity-driven visual grounding is conducted with SAM3 [6]. All textual and visual elements are embedded into a shared semantic space using `EVA-CLIP-8B` [68]. For **Multi-Granularity Graph Retrieval**, both dense similarity computation and Personalized PageRank (PPR) propagation are implemented on GPUs to accelerate large-scale inference. Hyperparameters are empirically tuned for each task, with detailed configurations and prompts provided in Appendix A.3. All experiments are conducted on two NVIDIA RTX 6000 Ada GPUs (48GB).

4.2 Graph Construction Efficiency

Construction Efficiency. To evaluate the efficiency of our hierarchical MMKG construction strategy, we compare the graph construction time and cost of MG^2-RAG with existing open-source graph-based MM-RAG baselines. For large benchmarks (E-VQA and InfoSeek), we restrict the knowledge base to a 5k subset because baseline approaches rely on expensive MLLM-driven triplet extraction, making full-scale graph construction computationally and financially infeasible. In contrast, MG^2-RAG remains efficient enough to construct graphs over the entire knowledge base, highlighting its scalability advantage.

Figure 3 summarizes the results. By bypassing the bottleneck of MLLM-based relation extraction and instead combining lightweight textual parsing with entity-driven visual grounding, MG^2-RAG achieves substantial efficiency gains across all evaluated datasets. The calculation method is detailed in Appendix A.2.

Table 1: Scalability of graph construction on E-VQA. We report the construction time under different knowledge-base sizes. 100k* denotes the setting with double GPUs.

Data Size (N)	1k	5k	10k	50k	100k	100k*
Time (Hours)	0.6	2.9	5.7	29.7	63.4	32.8

Table 2: Retrieval performance on E-VQA and InfoSeek. Retrieval modes include **T→V** (Text-to-Vision), **V→T** (Vision-to-Text), and **MM** (Multimodal retrieval). **Bold** and underlined denote the best and second-best results.

Model	Retrieval Modality	E-VQA					InfoSeek				
		R@1↑	R@5↑	R@10↑	R@20↑	R@50↑	R@1↑	R@5↑	R@10↑	R@20↑	R@50↑
Single Modality Retrieval											
Nomic-Embed [61]	T→T	4.7	9.1	12.1	15.9	23.0	0.2	1.0	1.5	3.0	5.6
	V→V	16.4	26.9	33.6	38.4	44.5	31.4	48.8	54.5	60.0	66.1
CLIP ViT-L/14 [65]	T→T	1.0	2.0	2.5	3.2	4.4	0.1	0.2	0.2	0.4	0.7
	V→V	17.8	31.0	36.4	42.6	50.3	32.7	52.3	59.6	66.2	74.0
EVA-CLIP-8B [68]	T→T	2.2	4.1	5.2	7.0	9.5	0.1	0.7	1.4	2.3	4.5
	V→V	30.0	45.0	50.0	55.0	59.9	47.2	65	70.2	74.4	79.6
Cross-Modality Retrieval											
CLIP ViT-L/14 [65]	T→V	1.6	3.3	4.8	7.0	10.9	0.2	0.6	1.1	2.3	5.3
	V→T	10.9	26.2	35.8	44.6	55.6	20.8	40.7	49.2	57.6	67.7
EVA-CLIP-8B [68]	T→V	1.8	4.6	6.3	8.5	12.9	0.2	0.4	0.8	2.0	4.9
	V→T	42.0	62.6	69.5	75.6	83.2	56.5	76.3	82.2	86.4	90.6
Multi-Modality Retrieval											
mKG-RAG [79]	MM	–	–	–	–	–	49.7	71.6	78.0	82.5	89.1
MG^2-RAG (Ours)	MM	**44.9**	**66.4**	**72.0**	**79.4**	**83.9**	**59.6**	**78.9**	**83.8**	**87.4**	**91.0**

– **Construction Time:** MG^2-RAG dramatically accelerates graph construction. Compared to the strongest baseline for each dataset, our framework achieves a maximum speedup of **77.1**× on ScienceQA and an average speedup of **43.3**× across all benchmarks (ranging from 22.2× to 77.1×).
– **Construction Cost:** The lightweight strategy also significantly reduces financial overhead. MG^2-RAG lowers construction costs by up to **45.2**× on ScienceQA, with an average cost reduction of **23.9**× across the five datasets.

These results demonstrate that our MMKG construction strategy effectively tackles a major scalability bottleneck of existing graph-based MM-RAG methods, enabling practical deployment without sacrificing downstream performance.

Scalability Results. We further evaluate the scalability of MG^2-RAG on E-VQA by increasing the knowledge base from 1k to 100k documents. As shown in Table 1, graph construction scales nearly linearly, with processing time increasing from **0.6** to **63.4 hours**. Under the largest setting, using two GPUs further reduces the construction time to **32.8 hours**, demonstrating that MG^2-RAG scales efficiently to large multimodal knowledge bases.

4.3 Multimodal Retrieval Performance

Table 2 reports retrieval results on E-VQA and InfoSeek, comparing MG^2-RAG with both cross-modal retrieval models and graph-based baselines. Our

Table 3: Knowledge-based VQA performance on InfoSeek and E-VQA. T and **V** denote Text and Vision modalities, respectively.

Model	LLM/MLLM	E-VQA		InfoSeek		
		Single-Hop↑	All↑	Unseen-Q↑	Unseen-E↑	All↑
Zero-Shot MLLMs						
InstructBLIP [14]	Flan-T5XL	11.90	12.00	8.90	7.40	8.10
LLaVA-v1.5 [49]	LLaMA-3.1-8B	16.00	16.90	8.30	8.90	7.80
Qwen2.5-VL-7B [13]	Qwen2.5-VL-7B	23.60	23.20	22.80	24.10	23.70
Gemini-3.1-Pro [20]	Gemini-3.1-Pro	29.36	27.90	29.92	22.52	25.70
GPT-5.2 [62]	GPT-5.2	44.19	40.30	32.68	26.54	29.29
Qwen3.5-27B [64]	Qwen3.5-27B	29.79	25.72	20.25	16.75	18.33
Retrieval-Augmented Models						
RORA-VLM [63]	Vicuna-7B	–	–	25.10	27.30	–
EchoSight [74] (V→T)	LLaMA-3.1-8B	52.96	47.23	30.00	30.70	30.40
EchoSight [74] (V→V)	LLaMA-3.1-8B	46.32	41.29	18.00	19.80	18.80
Graph Retrieval-Augmented Models						
mKG-RAG [79]	LLaMA-3.1-8B	–	–	32.90	31.30	32.10
MG^2-RAG(Ours)	LLaMA-3.1-8B	55.57	47.77	32.32	32.84	32.58
MG^2-RAG(Ours)	Qwen2.5-VL-7B	57.33	48.59	35.78	35.18	35.48
MG^2-RAG(Ours)	Qwen3.5-27B	61.65	52.19	37.36	38.40	37.87
MG^2-RAG(Ours)	GPT-5.2	<u>68.96</u>	**60.30**	<u>40.16</u>	<u>38.47</u>	<u>39.30</u>
VaLiK (5k) [51]	Qwen2.5-VL-7B	16.16	15.22	3.19	2.07	2.51
MMGraphRAG (5k) [69]	Qwen2.5-VL-7B	19.12	16.52	0.69	0.39	0.50
MG^2-RAG (5k) (Ours)	Qwen2.5-VL-7B	62.88	53.36	39.17	38.15	38.65
MG^2-RAG (5k) (Ours)	Qwen3.5-27B	**70.27**	<u>60.24</u>	**42.49**	**42.03**	**42.26**

framework consistently achieves the highest recall across all metrics (R@1–R@50). For instance, MG^2-RAG reaches R@1 scores of **44.9** and **59.6** on E-VQA and InfoSeek, respectively, as well as R@10 scores of **72.0** and **83.8**. These results surpass the strongest cross-modal baseline EVA-CLIP-8B ($V \rightarrow T$), which obtains 42.0 / 56.5 at R@1 and 69.5 / 82.2 at R@10.

Compared with the multimodal graph baseline mKG-RAG, our framework achieves a further +**9.9** improvement at R@1 on InfoSeek. These gains directly validate the effectiveness of our **multi-granularity graph retrieval mechanism**. By aggregating dense similarities onto unified multimodal nodes, MG^2-RAG bridges textual and visual evidence within a shared graph structure, allowing retrieval to capture cross-modal dependencies and structural relationships, thereby retrieving more accurate evidence than modality-isolated approaches.

4.4 Knowledge-based VQA Performance

Table 3 presents results on knowledge-based VQA, where models must reason over retrieved multimodal evidence. MG^2-RAG establishes a new state-of-the-art, consistently outperforming zero-shot MLLMs, standard RAG methods, and leading graph-based approaches. Notably, with the Qwen3.5-27B backbone, MG^2-RAG attains **60.24** on E-VQA (5k) and **42.26** on InfoSeek (5k). Moreover, MG^2-RAG also benefits strong closed-source models. For example, with GPT-5.2, the E-VQA score improves from **40.30** to **60.30**, showing that retrieval still provides useful knowledge even when the base MLLM is strong.

Table 4: Multimodal reasoning performance on ScienceQA. Results are reported across different domains and contexts: **NAT** (natural science), **SOC** (social science), **LAN** (language science), **TXT** (text context), **IMG** (image context), **NO** (no context), **G1-6** (grades 1-6), and **G7-12** (grades 7-12).

Model	LLM/MLLM	Subject			Context Modality			Grade		Avg.↑
		NAT↑	SOC↑	LAN↑	TXT↑	IMG↑	NO↑	G1-6↑	G7-12↑	
Human [53]	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
Zero-shot/Few-shot Models										
CoT [54]	GPT-4	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04	83.99
Chameleon [54]	ChatGPT	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53	79.93
Qwen2.5-VL-7B [3]	Qwen2.5-VL-7B	88.77	88.86	83.82	87.39	88.05	86.06	89.94	83.12	87.50
Gemini-3.1-Pro [20]	Gemini-3.1-Pro	93.53	94.12	90.81	92.98	94.25	90.27	94.03	90.93	92.90
GPT-5.2 [62]	GPT-5.2	97.60	93.05	95.96	97.73	95.35	96.17	96.54	95.88	96.30
Qwen3.5-27B [64]	Qwen3.5-27B	98.27	95.16	97.18	97.90	**96.43**	97.49	97.76	96.57	97.34
Graph Retrieval-Augmented Models										
MMKG [52]	Qwen2.5-7B	73.98	66.37	78.18	71.65	64.30	79.65	76.51	68.03	73.47
Visual Genome [37]	Qwen2.5-7B	76.78	67.04	78.09	74.05	66.19	79.72	78.08	69.68	75.08
VaLiK [51]	Qwen2.5-7B	84.15	75.14	87.64	82.99	73.18	89.69	84.40	80.95	83.16
VaLiK [51]	Qwen2.5-72B	85.61	75.93	90.27	84.40	74.17	92.33	85.79	82.98	84.77
MMGraphRAG [69]	Qwen2.5-VL-7B	81.08	68.62	80.09	79.52	68.96	83.69	80.87	73.43	78.21
MG^2-RAG(Ours)	Qwen2.5-7B	85.17	81.33	85.82	83.43	76.10	89.41	86.12	81.67	84.53
MG^2-RAG(Ours)	Qwen2.5-72B	88.37	81.21	91.27	87.00	78.28	94.01	88.84	85.43	87.62
MG^2-RAG(Ours)	Qwen2.5-VL-7B	90.81	87.96	87.09	89.93	86.61	88.92	90.64	86.75	89.25
MG^2-RAG(Ours)	GPT-5.2	97.78	94.65	97.06	97.93	96.24	97.05	97.33	96.43	97.00
MG^2-RAG(Ours)	Qwen3.5-27B	**98.45**	**95.28**	**98.73**	**98.44**	**96.43**	**98.68**	**98.42**	**96.84**	**97.85**

Two key advantages explain these improvements. First, compared with standard retrieval models like EchoSight [74], MG^2-RAG consistently yields better results, achieving +0.54 accuracy on E-VQA and +2.18 on InfoSeek. This improvement highlights the effectiveness of our **modality-preserving multimodal node fusion**, which retains both textual entities and visual objects as unified concepts. Second, MG^2-RAG outperforms existing graph-based approaches even when using the same backbone models. For instance, with the LLaMA-3.1-8B backbone, our framework already surpasses mKG-RAG by 0.48 on InfoSeek, and the advantage becomes larger with stronger models. This demonstrates that our **lightweight MMKG construction strategy** produces a cleaner and more informative graph topology than traditional MLLM-generated triplet graphs.

4.5 Multimodal Reasoning Performance

Table 4 reports results on ScienceQA, a widely used benchmark for evaluating multimodal reasoning across diverse scientific domains. MG^2-RAG achieves state-of-the-art performance, consistently surpassing both zero-shot/few-shot MLLMs and existing graph-based RAG models. In particular, when paired with the Qwen3.5-27B backbone, MG^2-RAG reaches an overall accuracy of **97.85**, substantially exceeding human performance (88.40) [53].

Two observations highlight the effectiveness of our framework. First, MG^2-RAG consistently outperforms prior graph-augmented models. Using the same Qwen2.5-VL-7B backbone, it improves over MMGraphRAG by 11.04 points in average accuracy and surpasses VaLiK by 1.37 points (7B) and 2.85 points (72B). These results demonstrate that our **multi-granularity graph representation**

Table 5: Multimodal classification performance on CrisisMMD. Evaluation includes three tasks: **BC** (binary informativeness classification), **MC** (fine-grained humanitarian category classification), and **MC-m** (merged-category classification).

Method	LLM/MLLM	CrisisMMD			
		BC↑	MC↑	MC-m↑	Avg.↑
Zero-Shot Models					
CLIP ViT-L/14 [65]	CLIP ViT-L/14	43.36	17.88	20.79	27.34
LLaVA-34B [50]	LLaVA-34B	56.44	25.15	25.07	35.55
BLIP-2 [44]	Flan-T5	61.29	40.86	40.72	47.62
Qwen2.5-VL-7B [3]	Qwen2.5-VL-7B	70.45	41.44	42.33	51.41
Gemini-3.1-Pro [20]	Gemini-3.1-Pro	71.30	49.36	49.79	56.82
GPT-5.2 [62]	GPT-5.2	71.00	<u>53.00</u>	<u>53.60</u>	<u>59.20</u>
Qwen3.5-27B [64]	Qwen3.5-27B	69.83	48.82	49.35	56.00
Graph Retrieval-Augmented Models					
LightRAG [23]	Qwen2.5-7B	67.49	45.11	45.94	52.85
VaLiK [51]	Qwen2.5-7B	68.90	50.02	50.69	56.54
VaLiK [51]	Qwen2.5-72B	68.89	49.78	49.31	55.99
MMGraphRAG [69]	Qwen2.5-VL-7B	68.40	44.03	44.66	50.94
MG^2-RAG(Ours)	Qwen2.5-VL-7B	71.03	46.45	47.25	54.91
MG^2-RAG(Ours)	Qwen2.5-7B	69.69	51.1	51.59	57.46
MG^2-RAG(Ours)	Qwen2.5-72B	**72.33**	50.34	50.83	57.83
MG^2-RAG(Ours)	Qwen3.5-27B	72.28	52.30	52.79	59.12
MG^2-RAG(Ours)	GPT-5.2	<u>72.30</u>	**56.10**	**56.70**	**61.70**

more effectively captures reasoning chains and structural relationships than conventional graph formulations. Second, MG^2-RAG improves the reasoning capability of strong base MLLMs. Compared with zero-shot inference, integrating MG^2-RAG yields average improvements of 1.75 points for Qwen2.5-VL-7B and 0.70 points for GPT-5.2. The gains on ScienceQA further show the robustness of our **hierarchical MMKG**: when fine-grained grounding is imperfect, chunk and image nodes still provide coarser contextual evidence for reasoning.

4.6 Multimodal Classification Performance

Table 5 presents results on CrisisMMD, evaluating multimodal classification in disaster scenarios. MG^2-RAG achieves state-of-the-art performance. In particular, the combination of MG^2-RAG and GPT-5.2 obtains the best overall accuracy of **61.70**, demonstrating strong performance in both binary relevance filtering (BC) and multi-class humanitarian classification (MC and MC-m).

Two insights emerge from the results. First, MG^2-RAG consistently surpasses existing graph-based retrieval frameworks. Using the Qwen2.5-VL-7B backbone, our model improves over MMGraphRAG by 3.97 points in average accuracy. With Qwen2.5-7B, it outperforms LightRAG by 4.61 points and VaLiK by 0.92 points, while maintaining a 1.84-point advantage over VaLiK with the 72B backbone. These improvements highlight the effectiveness of our **multi-granularity graph retrieval mechanism**, which captures complex cross-modal relationships within noisy real-world data. Second, MG^2-RAG substantially enhances the zero-shot capability of standard MLLMs. Compared with their

Table 6: Ablation study of MG^2-RAG. HGC, MNF, and GP denote Hierarchical Graph Construction, Multimodal Node Fusion, and Graph Propagation, respectively.

Method	E-VQA (5k) R@1/R@5↑	E-VQA (5k) BEM Score (All↑)	ScienceQA Acc. (Avg.↑)	CrisisMMD Acc. (Avg.↑)
MG^2-RAG	**57.8 / 83.1**	**60.24**	**97.85**	**59.12**
w/o HGC	40.7 / 61.5	51.72	97.48	56.70
w/o MNF	43.8 / 78.0	55.38	97.67	58.62
w/o GP	25.5 / 76.0	54.69	97.51	58.90

respective zero-shot baselines, our framework improves performance by 3.50 points for Qwen2.5-VL-7B and 2.50 points for GPT-5.2. This demonstrates that our **entity-driven visual grounding** helps resolve ambiguity in social-media content and improves decision reliability.

4.7 Ablation Study

Table 6 presents an ablation study of the three key components in MG^2-RAG: Hierarchical Graph Construction (**HGC**), Multimodal Node Fusion (**MNF**), and Graph Propagation (**GP**). Removing any component consistently degrades performance, demonstrating that the three designs complement each other to enable effective multimodal retrieval. Two observations are particularly noteworthy.

First, the **hierarchical graph structure** is essential for organizing multimodal evidence. Removing HGC collapses the hierarchical chunk-image-node representation into a flatter retrieval space, reducing E-VQA R@1/R@5 from 57.8/83.1 to 40.7/61.5. This confirms that the hierarchy effectively bridges coarse document context with fine-grained visual and textual evidence. Second, **cross-modal node alignment** and **graph-based relevance propagation** improve retrieval from complementary perspectives. Removing MNF decreases the E-VQA BEM score from 60.24 to 55.38, demonstrating the importance of aligning textual entities with grounded visual regions. In contrast, removing GP causes the largest drop in retrieval accuracy, reducing R@1 to 25.5, indicating that graph propagation is crucial for discovering structurally related evidence beyond initial dense retrieval. Together, these results show that MG^2-RAG benefits from hierarchy, multimodal alignment, and graph propagation in a complementary way.

4.8 Case Study

Figure 4 presents qualitative comparisons on knowledge-based VQA and multimodal retrieval. Baseline MLLMs (e.g., Qwen3.5-27B) frequently generate hallucinated or inaccurate answers due to the lack of external knowledge. For example, in Case (c), the baseline incorrectly predicts *Indonesia*. In contrast, MG^2-RAG produces accurate, grounded responses by **fusing textual entities and visual objects into unified multimodal nodes**, preserving fine-grained cross-modal evidence. This enables precise retrieval of details such as the wingspan in Case (b) and plant edibility in Case (d). Moreover, **multi-granularity graph retrieval** supports effective cross-modal reasoning. In Case (c), MG^2-RAG retrieves the key evidence "the pristine lake environment" and propagates it to

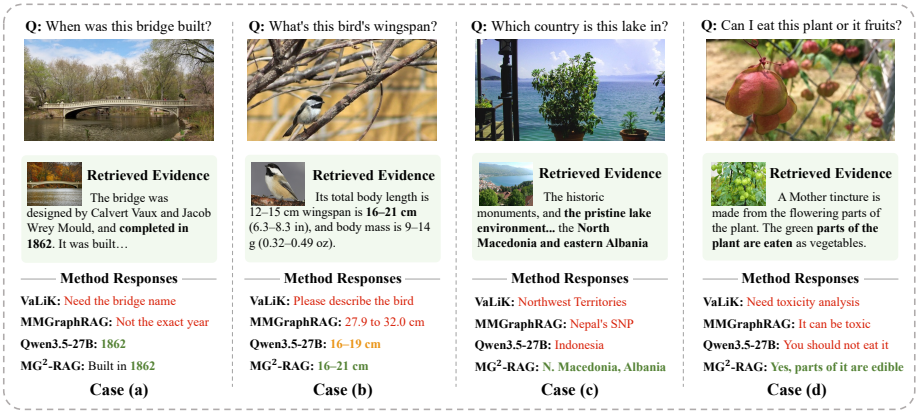

Fig. 4: Case studies on Knowledge-based VQA and Multimodal Retrieval. MG^2-RAG is compared with a baseline MLLM (Qwen3.5-27B) and graph-based approaches (VaLiK and MMGraphRAG). The **Retrieved Evidence** illustrates the multimodal evidence retrieved by MG^2-RAG for answer generation.

the corresponding image, correctly identifying the location as "North Macedonia, Albania." Additional qualitative examples for multimodal reasoning, classification, and representative failure cases are provided in Appendix B.

5 Conclusion

In this paper, we introduce **MG^2-RAG**, a lightweight Multi-Granularity Graph RAG framework that improves the reliability and efficiency of MM-RAG. We propose a hierarchical MMKG construction strategy that bypasses expensive MLLM-driven triplet extraction by combining lightweight textual parsing with entity-driven visual grounding, enabling efficient fusion of textual entities and visual objects into unified multimodal nodes. Building on this representation, we further design a multi-granularity graph retrieval mechanism that aggregates dense similarities and propagates relevance through the graph topology to support structured multi-hop reasoning. Extensive experiments across four multimodal tasks (i.e., retrieval, knowledge-based VQA, reasoning, and classification) show that MG^2-RAG consistently achieves state-of-the-art performance while maintaining high efficiency, providing an average 43.3× speedup and 23.9× cost reduction compared with existing graph-based baselines. Ultimately, MG^2-RAG provides a scalable, efficient, and robust solution for MM-RAG, enabling reliable deployment of MLLMs in complex, real-world knowledge-intensive applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62125201, U24B20174, and U25B6003, as well as the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123302).

References

1. Alam, F., Ofli, F., Imran, M.: CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In: Proceedings of the International AAAI Conference on Web and Social Media (AAAI). vol. 12 (2018), https://ojs.aaai.org/index.php/ICWSM/article/view/14983
2. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In: The Twelfth International Conference on Learning Representations (ICLR). vol. 2024, pp. 9112–9141 (2024), https://openreview.net/forum?id=hSyW5go0v8
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923 (2025), https://arxiv.org/abs/2502.13923
4. Bulian, J., Buck, C., Gajewski, W., Börschinger, B., Schuster, T.: Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 291–305 (2022), https://aclanthology.org/2022.emnlp-main.20/
5. Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1818–1826 (2024), https://openaccess.thecvf.com/content/CVPR2024W/MMFM/html/Caffagni_Wiki-LLaVA_Hierarchical_Retrieval-Augmented_Generation_for_Multimodal_LLMs_CVPRW_2024_paper.html
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment Anything with Concepts. arXiv preprint arXiv:2511.16719 (2025), https://arxiv.org/abs/2511.16719
7. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading Wikipedia to Answer Open-Domain Questions. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1870–1879 (2017), https://aclanthology.org/P17-1171/
8. Chen, L., Tong, P., Jin, Z., Sun, Y., Ye, J., Xiong, H.: Plan-on-Graph: Self-Correcting Adaptive Planning of Large Language Model on Knowledge Graphs. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 37, pp. 37665–37691 (2024), https://openreview.net/forum?id=CwCUEr6wO5
9. Chen, W., Hu, H., Chen, X., Verga, P., Cohen, W.: MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5558–5570 (2022), https://aclanthology.org/2022.emnlp-main.375/
10. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can Pre-trained Vision and Language Models Answer Visual Information-seeking Questions? In: Proceedings of the 2023 Conference on Empirical Methods in Natural

Language Processing (EMNLP). pp. 14948–14968 (2023), https://aclanthology.org/2023.emnlp-main.925/
11. Cheng, Q., Li, X., Li, S., Zhu, Q., Yin, Z., Shao, Y., Li, L., Sun, T., Yan, H., Qiu, X.: Unified Active Retrieval for Retrieval Augmented Generation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 17153–17166 (2024), https://aclanthology.org/2024.findings-emnlp.999/
12. Cocchi, F., Moratelli, N., Cornia, M., Baraldi, L., Cucchiara, R.: Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9199–9209 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Cocchi_Augmenting_Multimodal_LLMs_with_Self-Reflective_Tokens_for_Knowledge-based_Visual_Question_CVPR_2025_paper.html
13. Compagnoni, A., Morini, M., Sarto, S., Cocchi, F., Caffagni, D., Cornia, M., Baraldi, L., Cucchiara, R.: ReAG: Reasoning-Augmented Generation for Knowledge-based Visual Question Answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11901–11911 (2026), https://openaccess.thecvf.com/content/CVPR2026/html/Compagnoni_ReAG_Reasoning-Augmented_Generation_for_Knowledge-based_Visual_Question_Answering_CVPR_2026_paper.html
14. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). vol. 36, pp. 49250–49267 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html
15. Dong, J., An, S., Yu, Y., Zhang, Q.W., Luo, L., Huang, X., Wu, Y., Yin, D., Sun, X.: Youtu-GraphRAG: Vertically Unified Agents for Graph Retrieval-Augmented Complex Reasoning. In: The Fourteenth International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=yCtgZ2G39E
16. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R.O., Larson, J.: From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv preprint arXiv:2404.16130 (2024), https://arxiv.org/abs/2404.16130
17. Edge, D., Trinh, H., Larson, J.: LazyGraphRAG: Setting a New Standard for Quality and Cost (2024), https://www.microsoft.com/en-us/research/blog/lazygraphrag-setting-a-new-standard-for-quality-and-cost/
18. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., HUDELOT, C., Colombo, P.: ColPali: Efficient Document Retrieval with Vision Language Models. In: The Thirteenth International Conference on Learning Representations (ICLR). vol. 2025, pp. 61424–61449 (2025), https://openreview.net/forum?id=ogjBpZ8uSi
19. Gao, J., Li, L., Ji, K., Li, W., Lian, Y., Fu, Y., Dai, B.: SmartRAG: Jointly Learn RAG-Related Tasks From the Environment Feedback. In: The Thirteenth International Conference on Learning Representations (ICLR). vol. 2025, pp. 53117–53140 (2025), https://openreview.net/forum?id=OCd3cffulp
20. Google DeepMind Team: Gemini 3.1 Pro: Best for Complex Tasks and Bringing Creative Concepts to Life (2026), https://deepmind.google/models/gemini/pro/
21. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition

(CVPR). pp. 6904–6913 (2017), https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html
22. Guan, X., Zeng, J., Meng, F., Xin, C., Lu, Y., Lin, H., Han, X., Sun, L., Zhou, J.: DeepRAG: Thinking to Retrieval Step by Step for Large Language Models. In: The 14th International Conference on Learning Representations (ICLR) (2026), https://openreview.net/pdf?id=VI2YaggHIF
23. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: LightRAG: Simple and Fast Retrieval-Augmented Generation. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 10746–10761 (2025), https://aclanthology.org/2025.findings-emnlp.568/
24. Gutiérrez, B.J., Shu, Y., Gu, Y., Yasunaga, M., Su, Y.: HippoRAG: neurobiologically inspired long-term memory for large language models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS). vol. 37, pp. 59532–59569 (2024), https://openreview.net/forum?id=hkujvAPVsg
25. Gutiérrez, B.J., Shu, Y., Qi, W., Zhou, S., Su, Y.: From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In: International Conference on Machine Learning (ICML). pp. 21497–21515 (2025), https://openreview.net/forum?id=LWH8yn4HS2
26. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval Augmented Language Model Pre-Training. In: Proceedings of the 37th International Conference on Machine Learning (ICML). pp. 3929–3938 (2020), https://proceedings.mlr.press/v119/guu20a.html
27. Hao, J., Liu, H., Xiao, X., Huang, Q., Yu, J.: Uni-X: Mitigating Modality Conflict with a Two-End-Separated Architecture for Unified Multimodal Models. In: The Fourteenth International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=IJLIYpCkwz
28. Haveliwala, T.H.: Topic-sensitive pagerank. In: Proceedings of the 11th international conference on World Wide Web (WWW). pp. 517–526 (2002), https://dl.acm.org/doi/abs/10.1145/511446.511513
29. He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., Bresson, X., Hooi, B.: G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 37, pp. 132876–132907 (2024), https://openreview.net/forum?id=MPJ3oXtTZl
30. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-Strength Natural Language Processing in Python (2020), https://spacy.io/
31. Hsu, S., Khattab, O., Finn, C., Sharma, A.: Grounding by Trying: LLMs with Reinforcement Learning-Enhanced Retrieval. In: The Thirteenth International Conference on Learning Representations (ICLR). vol. 2025, pp. 26136–26152 (2025), https://openreview.net/forum?id=BPAZ6yW3K7
32. Hu, C.W., Wang, Y., Xing, S., Chen, C.J., Feng, S., Rossi, R., Tu, Z.: mRAG: Elucidating the Design Space of Multi-modal Retrieval-Augmented Generation. arXiv preprint arXiv:2505.24073 (2025), https://arxiv.org/abs/2505.24073
33. Hu, Z., Iscen, A., Sun, C., Wang, Z., Chang, K.W., Sun, Y., Schmid, C., Ross, D.A., Fathi, A.: REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23369–23379 (2023), https://openaccess.thecvf.com/content/CVPR2023/html/Hu_REVEAL_Retrieval-Augmented_Visual-Language_Pre-Training_With_Multi-Source_Multimodal_Knowledge_Memory_CVPR_2023_paper.html

34. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems (TOIS) **43**(2), 1–55 (2025), https://dl.acm.org/doi/full/10.1145/3703155
35. Huang, Y., Zhang, S., Xiao, X.: KET-RAG: A Cost-Efficient Multi-Granular Indexing Framework for Graph-RAG. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). pp. 1003–1012 (2025), https://dl.acm.org/doi/abs/10.1145/3711896.3737012
36. Jin, B., Xie, C., Zhang, J., Roy, K.K., Zhang, Y., Li, Z., Li, R., Tang, X., Wang, S., Meng, Y., Han, J.: Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 163–184 (2024), https://aclanthology.org/2024.findings-acl.11/
37. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Li, F.F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (2017), https://link.springer.com/article/10.1007/S11263-016-0981-7
38. Lazaridou, A., Gribovskaya, E., Stokowiec, W., Grigorev, N.: Internet-augmented language models through few-shot prompting for open-domain question answering. arXiv preprint arXiv:2203.05115 (2022), https://arxiv.org/abs/2203.05115
39. Lee, J., Wang, Y., Li, J., Zhang, M.: Multimodal Reasoning with Multimodal Knowledge Graph. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 10767–10782 (2024), https://aclanthology.org/2024.acl-long.579/
40. Lee, K., Chang, M.W., Toutanova, K.: Latent Retrieval for Weakly Supervised Open Domain Question Answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 6086–6096 (2019), https://aclanthology.org/P19-1612/
41. Lee, M., An, S., Kim, M.S.: PlanRAG: A Plan-then-Retrieval Augmented Generation for Generative Large Language Models as Decision Makers. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). pp. 6537–6555 (2024), https://aclanthology.org/2024.naacl-long.364/
42. Lee, Z., Cao, S., Liu, J., Zhang, J., Liu, W., Che, X., Hou, L., Li, J.: ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation. arXiv preprint arXiv:2503.21729 (2025), https://arxiv.org/abs/2503.21729
43. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). pp. 9459–9474 (2020), https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
44. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 19730–19742 (2023), https://proceedings.mlr.press/v202/li23q

45. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: Proceedings of the 39th International Conference on Machine Learning (ICML). pp. 12888–12900 (2022), https://proceedings.mlr.press/v162/li22n.html
46. Li, M., Miao, S., Li, P.: Simple is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation. In: International Conference on Learning Representations (ICLR). vol. 2025, pp. 6061–6089 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/11e1900e680f5fe1893a8e27362dbe2c-Paper-Conference.pdf
47. Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., Dou, Z.: Search-o1: Agentic Search-Enhanced Large Reasoning Models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5420–5438 (2025), https://aclanthology.org/2025.emnlp-main.276/
48. Lim, Q.Z., Lee, C.P., Lim, K.M., Samingan, A.K.: UniRaG: Unification, Retrieval, and Generation for Multimodal Question Answering With Pre-Trained Language Models. IEEE Access **12**, 71505–71519 (2024), https://doi.org/10.1109/ACCESS.2024.3403101
49. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html
50. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). vol. 36, pp. 34892–34916 (2023), https://openreview.net/forum?id=w0H2xGHlkw
51. Liu, J., Meng, S., Gao, Y., Mao, S., Cai, P., Yan, G., Chen, Y., Bian, Z., Wang, D., Shi, B.: Aligning Vision to Language: Annotation-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 981–992 (2025), https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Aligning_Vision_to_Language_Annotation-Free_Multimodal_Knowledge_Graph_Construction_for_ICCV_2025_paper
52. Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., Rosenblum, D.S.: MMKG: Multi-modal Knowledge Graphs. In: European Semantic Web Conference (ESWC). pp. 459–474 (2019), https://link.springer.com/chapter/10.1007/978-3-030-21348-0_30
53. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). vol. 35, pp. 2507–2521 (2022), https://openreview.net/forum?id=HjwK-Tc_Bc
54. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.W., Wu, Y.N., Zhu, S.C., Gao, J.: Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). vol. 36, pp. 43447–43478 (2023), https://openreview.net/forum?id=HtqnVSCj3q
55. Luo, L., Li, Y.F., Haffari, G., Pan, S.: Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In: The Twelfth International Conference on Learning Representations (ICLR). vol. 2024, pp. 14400–14423 (2024), https://openreview.net/forum?id=ZGNWW7xZ6Q

56. Luo, L., Zhao, Z., Haffari, G., Phung, D., Gong, C., Pan, S.: GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 38, pp. 36371–36405 (2025), https://openreview.net/forum?id=0QNmAvQQqj
57. Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., Ji, R.: Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 38, pp. 168008–168033 (2025), https://openreview.net/forum?id=QaZxGWlbgO
58. Ma, S., Xu, C., Jiang, X., Li, M., Qu, H., Yang, C., Mao, J., Guo, J.: Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation. In: International Conference on Learning Representations (ICLR). vol. 2025, pp. 52782–52806 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/830b1abc6d2da85f23d41169fa44d185-Paper-Conference.pdf
59. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: Encyclopedic VQA: Visual Questions about Detailed Properties of Fine-grained Categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3113–3124 (2023), https://openaccess.thecvf.com/content/ICCV2023/html/Mensink_Encyclopedic_VQA_Visual_Questions_About_Detailed_Properties_of_Fine-Grained_Categories_ICCV_2023_paper.html
60. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: PlotQA: Reasoning over Scientific Plots. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1527–1536 (2020), https://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html
61. Nussbaum, Z., Morris, J.X., Mulyar, A., Duderstadt, B.: Nomic Embed: Training a Reproducible Long Context Text Embedder. Transactions on Machine Learning Research (TMLR) (2025), https://openreview.net/forum?id=IPmzyQSiQE
62. OpenAI Team: Introducing GPT-5.2:The Most Advanced Frontier Model for Professional Work and Long-running Agents. (2026), https://openai.com/index/introducing-gpt-5-2/
63. Qi, J., Xu, Z., Shao, R., Chen, Y., Di, J., Cheng, Y., Wang, Q., Huang, L.: RoRA-VLM: Robust Retrieval-Augmented Vision Language Models. arXiv preprint arXiv:2410.08876 (2024), https://arxiv.org/abs/2410.08876
64. Qwen Team: Qwen3.5: Towards Native Multimodal Agents (2026), https://qwen.ai/blog?id=qwen3.5
65. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763 (2021), https://proceedings.mlr.press/v139/radford21a
66. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D.: RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In: The Twelfth International Conference on Learning Representations (ICLR). vol. 2024, pp. 32628–32649 (2024), https://openreview.net/forum?id=GN921JHCRw
67. Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., Ni, L., Shum, H.Y., Guo, J.: Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In: The Twelfth International Conference on Learning

Representations (ICLR). vol. 2024, pp. 3868–3898 (2024), https://openreview.net/forum?id=nnVO1PvbTv
68. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. arXiv preprint arXiv:2402.04252 (2024), https://arxiv.org/abs/2402.04252
69. Wan, X., Yu, H.: MMGraphRAG: Bridging Vision and Language with Interpretable Multimodal Knowledge Graphs. arXiv preprint arXiv:2507.20804 (2025), https://arxiv.org/abs/2507.20804
70. Wang, L., Chen, H., Yang, N., Huang, X., Dou, Z., Wei, F.: Chain-of-Retrieval Augmented Generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS). pp. 59888–59915 (2025), https://openreview.net/forum?id=gUPGGCM4WH
71. Wang, S., Fang, Y., Zhou, Y., Liu, X., Ma, Y.: ArchRAG: Attributed Community-based Hierarchical Retrieval-Augmented Generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 15868–15876 (2026), https://ojs.aaai.org/index.php/AAAI/article/view/38619
72. Wasserman, N., Pony, R., Naparstek, O., Goldfarb, A.R., Schwartz, E., Barzelay, U., Karlinsky, L.: Real-mm-rag: A real-world multi-modal retrieval benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 31660–31683 (2025), https://aclanthology.org/2025.acl-long.1528/
73. Wu, J., Zhu, J., Liu, Y., Xu, M., Jin, Y.: Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 28489–28503 (2025), https://aclanthology.org/2025.acl-long.1383/
74. Yan, Y., Xie, W.: EchoSight: Advancing Visual-Language Models with Wiki Knowledge. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 1538–1551 (2024), https://aclanthology.org/2024.findings-emnlp.83/
75. Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., Yih, W.t.: Retrieval-Augmented Multimodal Language Modeling pp. 39755–39769 (2022), https://proceedings.mlr.press/v202/yasunaga23a.html
76. You, X., Huang, Q., Li, L., Chang, X., Yu, J.: Cut to the Chase: Training-free Multimodal Summarization via Chain-of-Events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26219–26229 (2026), https://openaccess.thecvf.com/content/CVPR2026/html/You_Cut_to_the_Chase_Training-free_Multimodal_Summarization_via_Chain-of-Events_CVPR_2026_paper.html
77. You, X., Huang, Q., Li, L., Zhang, C., Liu, X., Zhang, M., Yu, J.: Knowledge Completes the Vision: A Multimodal Entity-aware Retrieval-Augmented Generation Framework for News Image Captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 40, pp. 12108–12116 (2026), https://ojs.aaai.org/index.php/AAAI/article/view/38200
78. Yu, J., Liu, Y., Gu, J., Torr, P., Zhou, D.: Can Knowledge-Graph-based Retrieval Augmented Generation Really Retrieve What You Need? In: Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 38, pp. 95653–95682 (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/89d0d5c2f720921df93bbb8fef514571-Paper-Conference.pdf
79. Yuan, X., Ning, L., Fan, W., Li, Q.: mKG-RAG: Multimodal Knowledge Graph-Enhanced RAG for Visual Question Answering. In: Proceedings of the 49th Inter-

national ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2026), https://doi.org/10.1145/3805712.3809680
80. Zhang, J., Zhang, X., Yu, J., Tang, J., Tang, J., Li, C., Chen, H.: Subgraph Retrieval Enhanced Model for Multi-hop Knowledge Base Question Answering. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL). pp. 5773–5784 (2022), https://aclanthology.org/2022.acl-long.396/
81. Zhang, R., Liu, C., Su, Y., Li, R., Huang, X., Li, X., Yu, P.S.: A comprehensive survey on multimodal RAG: all combinations of modalities as input and output. Authorea Preprints (2025), https://www.techrxiv.org/doi/full/10.36227/techrxiv.176341513.38473003
82. Zhang, T., Zhang, Z., Ma, Z., Chen, Y., Qi, Z., Yuan, C., Li, B., Pu, J., Zhao, Y., Xie, Z., Ma, J., Shan, Y., Hu, W.: mR^{2}AG: Multimodal Retrieval-Reflection-Augmented Generation for Knowledge-Based VQA. arXiv preprint arXiv:2502.01113 (2024), https://arxiv.org/abs/2411.15041
83. Zhang, Z., Feng, Y., Zhang, M.: LevelRAG: Enhancing Retrieval-Augmented Generation with Multi-hop Logic Planning over Rewriting Augmented Searchers. arXiv preprint arXiv:2502.18139 (2025), https://arxiv.org/abs/2502.18139
84. Zhao, D., Huang, Q., Yan, D., Sun, Y., Yu, J.: Partially shared concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 40, pp. 13117–13125 (2026), https://ojs.aaai.org/index.php/AAAI/article/view/38312
85. Zhao, Y., Zhu, J., Guo, Y., He, K., Li, X.: E^{2}GraphRAG: Streamlining Graph-based RAG for High Efficiency and Effectiveness. arXiv preprint arXiv:2505.24226 (2025), https://arxiv.org/abs/2505.24226
86. Zhou, Y., Zhang, T., Xu, S., Chen, S., Zhou, Q., Tong, Y., Ji, S., Zhang, J., Qi, L., Li, X.: Are They the Same? Exploring Visual Correspondence Shortcomings of Multimodal LLMs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17663–17674 (2025), https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_Are_They_the_Same_Exploring_Visual_Correspondence_Shortcomings_of_Multimodal_ICCV_2025_paper.html
87. Zhuang, L., Chen, S., Xiao, Y., Zhou, H., Zhang, Y., Chen, H., Zhang, Q., Huang, X.: LinearRAG: Linear Graph Retrieval Augmented Generation on Large-scale Corpora. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=mCtfkypdm6