M^{3}Prune: Hierarchical Collaborative Pruning for Efficient Multi-Modal Multi-Agent Retrieval-Augmented Generation

Taolin Zhang**
tlzhang@hfut.edu.cn
Hefei University of Technology
Hefei, China

Weizi Shao*
East China Normal University
Shanghai, China
51265901007@stu.ecnu.edu.cn

Zijie Zhou
China University of Petroleum
Beijing, China
zjzhouzh@gmail.com

Chen Chen
Guangdong University of Finance and Economics
Guangdong, China
allen821@student.gdufe.edu.cn

Daiyang Yu
Hefei University of Technology
Hefei, China
2023216618@mail.hfut.edu.cn

Tingyuan Hu
East China Normal University
Shanghai, China
10245102409@stu.ecnu.edu.cn

Chengyu Wang✉
Alibaba Group
Hangzhou, China
chengyu.wcy@alibaba-inc.com

Xiaofeng He
East China Normal University
Shanghai, China
hexf@cs.ecnu.edu.cn

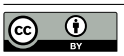
Abstract

Recent advances in multi-modal retrieval-augmented generation (mRAG), which augments multi-modal large language models (MLLMs) with external knowledge, have shown that collective intelligence from multiple agents can outperform a single model through effective communication. Despite their strong performance, existing multi-agent systems incur substantial token overhead and computational cost, posing challenges for large-scale deployment. To address these issues, we propose a **M**ulti-**M**odal **M**ulti-agent hierarchical communication graph **PRUNING** framework, termed **M^{3}Prune**. M^{3}Prune eliminates redundant communication edges both across and within modalities, improving the trade-off between task performance and token overhead. Specifically, M^{3}Prune first performs intra-modal graph sparsification in the textual and visual modalities to identify task-critical communication links. It then constructs an inter-modal communication graph and sparsifies cross-modal connections while encouraging consistent cross-modal reasoning through a modality alignment score. Finally, it progressively prunes redundant edges to obtain an efficient hierarchical topology. Extensive experiments on both general-domain and domain-specific mRAG benchmarks show that M^{3}Prune consistently outperforms single-agent and strong multi-agent mRAG systems while significantly improving token efficiency. The code and data are available at https://github.com/ztl-35/M3Prune.

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**.

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836330

Keywords

Multi-Modal Retrieval-Augmented Generation; Multi-Modal Multi-Agent Systems; Graph Optimization

ACM Reference Format:
Taolin Zhang, Weizi Shao, Zijie Zhou, Chen Chen, Daiyang Yu, Tingyuan Hu, Chengyu Wang, and Xiaofeng He. 2026. M^{3}Prune: Hierarchical Collaborative Pruning for Efficient Multi-Modal Multi-Agent Retrieval-Augmented Generation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836330

1 Introduction

While RAG has achieved significant success in the textual domain [8, 18, 50], extending it to the multi-modal setting remains challenging [15]. Traditional methods typically rely on a cascaded pipeline in which a multi-modal retriever extracts relevant evidence, which is then forwarded to LLMs for answer synthesis [13, 16, 33]. This segmented approach inherently suffers from global semantic misalignment [17, 21].

The emergence of multi-modal large language models (MLLMs) presents a transformative opportunity, as these models exhibit strong capabilities for modeling vision–language correlations [3, 11, 36]. This capability enables MLLMs to serve not only as generators but also as integrative engines for multi-modal retrieval-augmented generation (mRAG). By jointly processing retrieved multi-modal evidence and queries within a unified reasoning backbone, MLLMs can synthesize information and derive coherent conclusions from retrieved knowledge [28, 49, 53]. Nonetheless, when confronted with complex and multi-faceted multi-modal queries that require diverse expertise or deliberative reasoning, a single agent often reaches its performance limit [9, 30, 45].

Recent studies have explored multi-agent systems built upon MLLMs, where multiple specialized agents communicate through structured graph topologies to distribute reasoning workloads for complex multi-modal tasks [29, 31, 47]. However, such performance

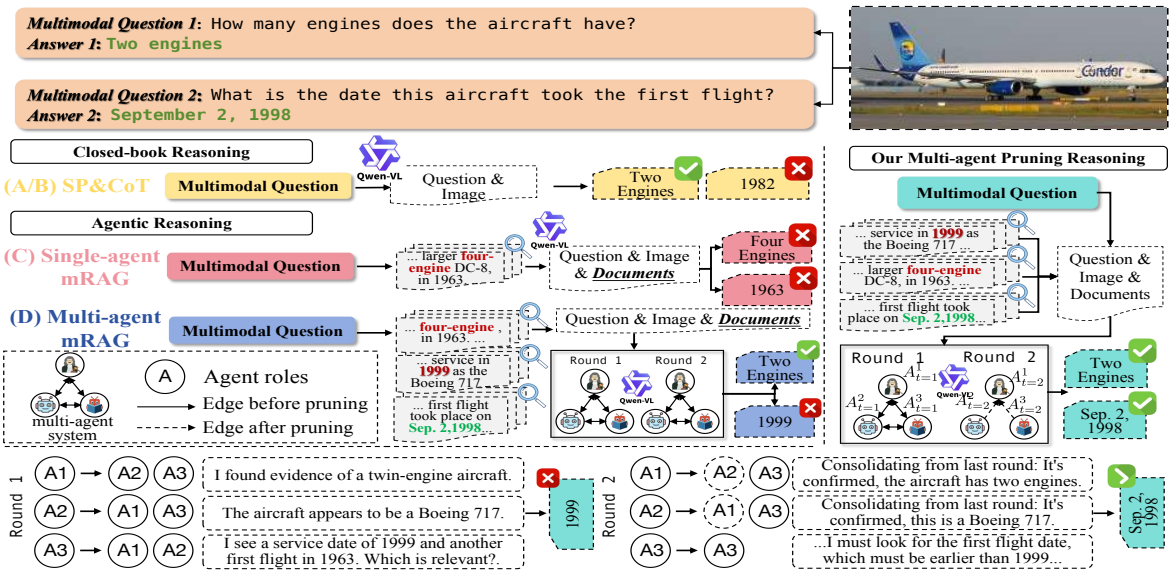


Figure 1: Comparison of our approach with existing methods. (1) `Closed-book Reasoning` does not utilize external knowledge. (2) `Single-agent mRAG` leverages an end-to-end MLLM coupled with a retriever to answer all questions. (3) `Multi-agent mRAG` constructs a fixed communication topology to obtain collaborative answers, without considering communication efficiency. (4) Our `Multi-agent Pruning` method prunes unnecessary communication edges, improving answer quality and token efficiency.

gains often come at the cost of substantial computational and token overhead. We identify the root cause as `communication redundancy`, which introduces inefficiency and can even undermine the accuracy of the final response [15, 48]. This redundancy is especially pronounced in multi-modal settings, where the information required to process textual and visual cues differs. Yet existing methods frequently adopt uniform communication strategies across modalities [19, 24, 40]. As illustrated in Fig. 1, in the "first flight date" example, the date-expert agent is overwhelmed by irrelevant messages about engine counts and model variants. Such noise hinders evidence aggregation and increases the likelihood of an incorrect answer.

In this paper, we introduce **M^{3}Prune**, a framework for hierarchical communication-graph pruning in multi-agent mRAG systems. M^{3}Prune learns task-adaptive communication topologies by assigning learnable softened adjacency matrices to spatio-temporal intra- and inter-modal connections, optimizing them for task utility under sparsity regularization, and progressively pruning edges with low importance. The core components of M^{3}Prune are:

Intra-Modal Graph Sparsification: In mRAG tasks, agents assigned different roles may produce divergent responses to the same query [46, 52]. Consequently, both cooperation and conflict may arise among agents within the same modality, potentially limiting mutual improvement. To address this issue, we design a spatio-temporal message-passing scheme that facilitates effective response exchange within the visual and textual modalities, respectively. We learn softened adjacency matrices that quantify the importance of intra-modal communication links based on their contribution to the final response. Edges that contribute little are sparsified, thereby focusing communication on key agents.

Inter-Modal Graph Sparsification: Given the heterogeneous granularity of information across modalities, it is crucial to integrate key clues from multiple modalities before generating the final response [2, 6, 20, 38]. Accordingly, we construct an inter-modal spatio-temporal communication graph to enable robust cross-modal collaboration and mitigate inconsistent responses. We initialize inter-modal connections and learn their softened adjacency matrices to assess the importance of cross-modal communication. In addition to optimizing task utility and sparsity regularization, we introduce a modality alignment score to encourage consistent semantic understanding across modalities during inter-modal sparsification. Finally, after learning sparsified intra- and inter-modal graphs, we progressively prune redundant edges to obtain an efficient hierarchical topology.

In experiments, we evaluate M^{3}Prune against strong baselines in zero-shot, single-agent, and multi-agent settings on several mRAG benchmarks: Vidoseek [39], MultimodalQA [35], and ScienceQA [25], covering both general and domain-specific tasks. The results show that M^{3}Prune achieves state-of-the-art performance, improving accuracy by **9.4%** and token efficiency by **23.8%** compared with strong multi-agent multi-modal baselines.

2 Related Work

2.1 Multi-Modal RAG

Prior research on mRAG can be broadly grouped into two themes: (1) **Modality-Adaptive Retrieval Strategies.** Early mRAG systems typically employed static retrieval methods, often incurring unnecessary computational overhead [10, 25, 27]. More recent work introduces dynamic, query-aware retrieval strategies that determine not only whether to retrieve but also which modality to retrieve from. For example, EchoSight [44] and RoRA-VLM [32] perform wiki-article retrieval using visual-only information, followed by re-ranking with respect to the combined text–visual query. (2) **Retrieval-Aware Pre-Training and Fine-Tuning.** CoRe-MMRAG [37]

addresses inconsistencies across knowledge sources through a four-stage progressive multi-modal retrieval pipeline. LLaVA-mR^{2}AG [51] introduces a fine-tuned adaptive retriever that derives answers through two reflection operations. Wiki-LLaVA [5] employs a hierarchical retrieval pipeline to integrate external knowledge from multi-modal documents. While these methods rely on a single MLLM backbone, they largely overlook the potential benefits of collective intelligence and multi-agent collaboration for mRAG [15].

2.2 Multi-Agent Systems for mRAG

This line of work can be broadly categorized into the following directions: (1) **General Agentic Collaboration.** A key shift has been the transition from static retrieval-and-generation pipelines to agentic designs, transforming passive systems into active decision-making frameworks [7, 23, 46]. ViDoRAG [39] introduces an iterative agent workflow that combines exploration, summarization, and reflection to enable scalable test-time reasoning in mRAG. OmniSearch [19] designs a self-adaptive planning agent that emulates human behavior by dynamically decomposing complex multi-modal queries into sub-questions with adaptive retrieval steps. E-agent [40] proposes a plan-then-execute architecture that combines a dynamic mRAG planner with a tool executor. (2) **Evolving Collaborative Architectures.** As mRAG tasks grow more complex, general multi-agent systems encounter increasing limitations, motivating research on specialized multi-modal agent collaboration [11, 34]. HM-RAG [24] employs a hierarchical agent structure, including a decomposer agent to break down multi-intent queries and a decision agent to synthesize the final answer. MuaLLM [1] integrates a hybrid mRAG framework with iterative reasoning, goal setting, and multi-step retrieval using an adaptive vector database. Although multi-agent collaboration yields considerable improvements on complex mRAG tasks, redundant and inefficient agent interactions remain a bottleneck, increasing both computational and token costs [15, 48]. Hence, we propose a multi-modal multi-agent edge-pruning framework that preserves task performance while reducing token overhead.

3 Basic Notations and Task Definition

Basic Notations. Our framework consists of intra- and inter-modal graph pruning modules. We represent each modality-specific communication structure as an augmented graph that encodes both connectivity and agent memory states. The intra-modal module comprises a textual graph, $\mathcal{G}_{txt}^{\text{intra}} = (\mathcal{V}_{txt}, \mathcal{E}_{txt}^{\mathcal{T}}, \mathcal{E}_{txt}^{\mathcal{S}}, \mathcal{S}_{txt})$, and a visual graph, $\mathcal{G}_{vis}^{\text{intra}} = (\mathcal{V}_{vis}, \mathcal{E}_{vis}^{\mathcal{T}}, \mathcal{E}_{vis}^{\mathcal{S}}, \mathcal{S}_{vis})$, where $\mathcal{V}_{txt} = \{v_i^{txt}\}$ and $\mathcal{V}_{vis} = \{v_i^{vis}\}$ denote the sets of textual and visual agent nodes, respectively. Each agent node is initialized with a corresponding textual or visual role. $\mathcal{E}_{txt}^{\mathcal{T}}$ and $\mathcal{E}_{txt}^{\mathcal{S}}$ denote the textual temporal and spatial edge sets, respectively,[1] and $\mathcal{S}_{txt}^{(t)} = \{s_i^{txt,(t)}\}$ represents the memory states of textual agents at round t. Correspondingly, $\mathcal{E}_{vis}^{\mathcal{T}}$ and $\mathcal{E}_{vis}^{\mathcal{S}}$ denote the visual temporal and spatial edge sets, and $\mathcal{S}_{vis}^{(t)} = \{s_i^{vis,(t)}\}$ represents the memory states of visual agents at round t. The memory update of each agent from round t to round $(t+1)$ is:

$$s_i^{m,(t+1)} = f_{\text{tr}}\big(s_i^{m,(t)}, \mathbf{q}, \mathbf{c}, I_{\mathcal{T}}^{m,(t+1)}, I_{\mathcal{S}}^{m,(t+1)}\big), \tag{1}$$

[1]Each agent is fully connected to all other agents except itself through spatial edges, and connected to all agents (including itself) through temporal edges across rounds.

where f_{tr} denotes the MLLM-based aggregation function that updates the agent memory at the current round. Here, $m \in \{\text{txt}, \text{vis}\}$ indexes the modality, and $\mathbf{q}$ and $\mathbf{c}$ denote the question and retrieved contexts, respectively. $I_{\mathcal{T}}^{m,(t+1)}$ and $I_{\mathcal{S}}^{m,(t+1)}$ denote the information aggregated from temporal and spatial neighbors, respectively, in modality m at round $(t+1)$.

For the inter-modal component, all textual and visual agents are incorporated into an inter-modal graph, $\mathcal{G}^{\text{inter}} = (\mathcal{V}, \mathcal{E}^{\mathcal{T}}, \mathcal{E}^{\mathcal{S}}, \mathcal{S})$, where the node set is defined as $\mathcal{V} = \mathcal{V}_{txt} \cup \mathcal{V}_{vis}$. The temporal edge set integrates both intra- and inter-modal connections: $\mathcal{E}^{\mathcal{T}} = \mathcal{E}_{txt}^{\mathcal{T}} \cup \mathcal{E}_{vis}^{\mathcal{T}} \cup \mathcal{E}_{vis\to txt}^{\mathcal{T}} \cup \mathcal{E}_{txt\to vis}^{\mathcal{T}}$. Similarly, the spatial edge set is defined as: $\mathcal{E}^{\mathcal{S}} = \mathcal{E}_{txt}^{\mathcal{S}} \cup \mathcal{E}_{vis}^{\mathcal{S}} \cup \mathcal{E}_{vis\to txt}^{\mathcal{S}} \cup \mathcal{E}_{txt\to vis}^{\mathcal{S}}$. The intra-modal edges in both temporal and spatial sets are inherited from the corresponding intra-modal graphs, while the inter-modal edges are initialized as fully connected. Finally, the joint agent memory is denoted as $\mathcal{S} = \mathcal{S}_{txt} \cup \mathcal{S}_{vis}$. The memory update mechanism is analogous to Eq. 1, except that the aggregated information $I_{\mathcal{T}}^{(t+1)}$ and $I_{\mathcal{S}}^{(t+1)}$ may also originate from inter-modal temporal and spatial edges. For example, $I_{\mathcal{T}}^{(t+1)}$ aggregates information from $I_{\mathcal{T}}^{txt,(t+1)}$ and $I_{\mathcal{T}}^{vis,(t+1)}$.

Task Definition. Given a question $\mathbf{q}$ and retrieved contexts $\mathbf{c}$, each agent in an intra-modal graph receives information solely through edges within the same modality, whereas agents in the inter-modal graph may also receive information from agents in the other modality. For each agent, its output response at round t is defined as $O_i^{(t)} = f_\theta\left(\mathbf{q}, \mathbf{c}, I_{\mathcal{T}}^{(t)}, I_{\mathcal{S}}^{(t)}\right)$, where f_θ denotes the MLLM-generated response conditioned on the question, retrieved contexts, and both temporal and spatial information available to the agent. Here, $I_{\mathcal{T}}^{(t)}$ and $I_{\mathcal{S}}^{(t)}$ refer to the temporal and spatial information aggregated for the current agent at round t from all available neighbors under the chosen communication graph (intra-modal or inter-modal). After multiple rounds of multi-agent interaction, a summary agent produces the final answer: $O_{\text{s}}^{(T)} = f_{\text{s}}\left(\mathbf{q}, \mathbf{c}, I_{\mathcal{T}}^{(T)}, I_{\mathcal{S}}^{(T)}\right)$, where T is the total number of rounds, f_{s} denotes the summary agent, and $O_{\text{s}}^{(T)}$ is the final answer.

However, a fixed communication topology inevitably introduces redundant information, motivating the need to prune superfluous edges [19, 40]. To formalize this issue, we define communication redundancy to guide the pruning process:

Definition 1 (Communication Redundancy). Given a multi-modal multi-agent communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}^{\mathcal{T}}, \mathcal{E}^{\mathcal{S}}, \mathcal{S})$ with a utility function $\phi(\cdot)$ that quantifies task performance, an edge $e \in \mathcal{E}^{\mathcal{T}} \cup \mathcal{E}^{\mathcal{S}}$ is considered redundant if removing it does not degrade task performance, i.e.,

$$\mathcal{G}^{sub} = (\mathcal{V}, \tilde{\mathcal{E}}^{\mathcal{T}}, \tilde{\mathcal{E}}^{\mathcal{S}}, \mathcal{S}), \quad \phi(\mathcal{G}^{sub}) \geq \phi(\mathcal{G}), \tag{2}$$

where $\tilde{\mathcal{E}}^{\mathcal{T}} = \begin{cases} \mathcal{E}^{\mathcal{T}} \setminus \{e\}, & e \in \mathcal{E}^{\mathcal{T}}, \\ \mathcal{E}^{\mathcal{T}}, & e \in \mathcal{E}^{\mathcal{S}}, \end{cases}$ and $\tilde{\mathcal{E}}^{\mathcal{S}} = \begin{cases} \mathcal{E}^{\mathcal{S}} \setminus \{e\}, & e \in \mathcal{E}^{\mathcal{S}}, \\ \mathcal{E}^{\mathcal{S}}, & e \in \mathcal{E}^{\mathcal{T}}. \end{cases}$

Specifically, the goal of inter-modal edge pruning is to construct a subgraph $\mathcal{G}^{sub}$ by removing redundant edges while preserving task performance [48]; this objective is further regularized by modality semantic alignment [2], as elaborated next.

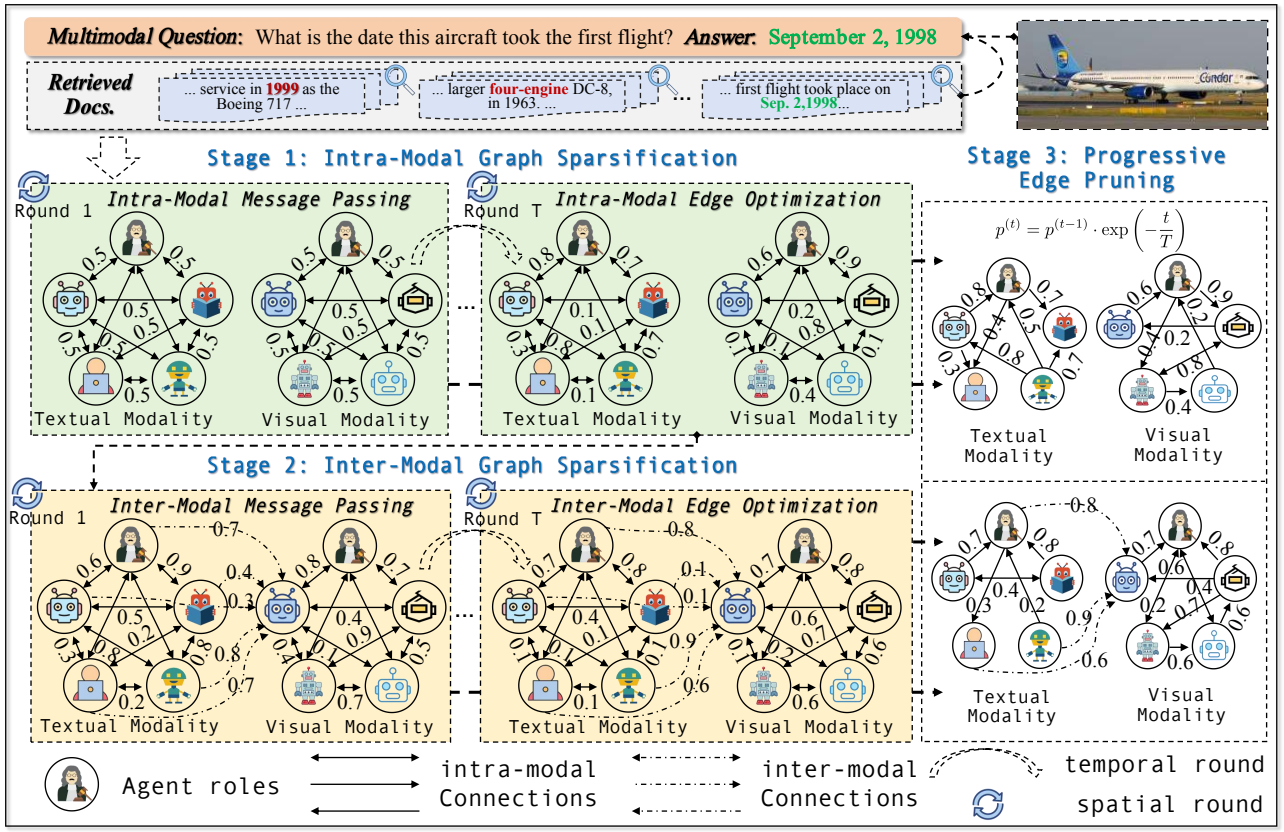


Figure 2: Overview. The key components include: (1) `Intra-modal Graph Sparsification`, which analyzes the input question using multiple agents for the textual and visual modalities, respectively; (2) `Inter-modal Graph Sparsification`, which integrates semantic information across modalities through interactions among multi-agent responses in the textual and visual domains; (3) `Progressive Edge Pruning`, which progressively prunes redundant edges *during training* to obtain a compact communication graph for inference.

4 Methodology

4.1 Intra-Modal Graph Sparsification

We design a spatio-temporal message-passing scheme to facilitate the effective exchange of task-relevant responses among agents within the visual and textual modalities. To improve the quality of the final response, we dynamically adjust edge weights and sparsify redundant connections, thereby focusing communication on critical interactions.

Edge Weight Initialization. Given the predefined spatial and temporal edges $\mathcal{E}_m^{\mathcal{S}}$ and $\mathcal{E}_m^{\mathcal{T}}$ for $m \in \{\text{txt}, \text{vis}\}$, we obtain the corresponding adjacency matrices $\mathbf{A}_m^{\mathcal{S}}$ and $\mathbf{A}_m^{\mathcal{T}}$ for the intra-modal communication graph. Based on $\mathbf{A}_m^{\mathcal{S}}$ and $\mathbf{A}_m^{\mathcal{T}}$, we initialize learnable softened adjacency matrices $\tilde{\mathbf{A}}_m^{\mathcal{S}}$ and $\tilde{\mathbf{A}}_m^{\mathcal{T}}$ as edge weights via Gumbel-Softmax [14] (Eq. 3), yielding values in $(0, 1)$. Taking $\tilde{\mathbf{A}}_m^{\mathcal{S}}$ as an example, each element $\tilde{\mathbf{A}}_m^{\mathcal{S}}[i, j]$ is computed as:

$$\tilde{\mathbf{A}}_m^{\mathcal{S}}[i,j] = \frac{\exp\left((\log(\mathbf{A}_m^{\mathcal{S}}[i,j] + \epsilon) + g_{ij})/\tau\right)}{\sum_{k=1}^{N_m} \exp\left((\log(\mathbf{A}_m^{\mathcal{S}}[i,k] + \epsilon) + g_{ik})/\tau\right)} \tag{3}$$

where $g_{ij} \sim \text{Gumbel}(0, 1)$, τ is the temperature, N_m is the number of agents in modality m, and ϵ is a small constant for numerical stability. $\tilde{\mathbf{A}}_m^{\mathcal{T}}$ is obtained similarly. To enable multi-round communication, we maintain round-specific softened adjacency matrices $\tilde{\mathbf{A}}_m^{(\mathcal{S},(t))}$ and $\tilde{\mathbf{A}}_m^{(\mathcal{T},(t))}$, which are updated at each round t.

Intra-Modal Message Passing. Within the intra-modal graph $\mathcal{G}^{\text{intra}}$, agent $v_i^{m,(t)}$ at round t receives the spatial message:

$$\mathcal{M}_{i,\text{intra}}^{(m,\mathcal{S},(t))} = \sum_{v_j^{m,(t)} \in \mathcal{N}_{\text{intra}}^{\mathcal{S}}(v_i^{m,(t)})} \mathcal{W}_m^{(\mathcal{S},(t))}[i,j] \cdot O(v_j^{m,(t)}), \tag{4}$$

where the edge attention weights are defined as

$$\mathcal{W}_m^{(\mathcal{S},(t))}[i,j] = \frac{\exp(\tilde{\mathbf{A}}_m^{(\mathcal{S},(t))}[i,j])}{\sum_{v_k^{m,(t)} \in \mathcal{N}_{\text{intra}}^{\mathcal{S}}(v_i^{m,(t)})} \exp(\tilde{\mathbf{A}}_m^{(\mathcal{S},(t))}[i,k])}. \tag{5}$$

Here, $\mathcal{N}_{\text{intra}}^{\mathcal{S}}(v_i^{m,(t)})$ denotes the set of spatial neighbors of $v_i^{m,(t)}$, and $O(v_j^{m,(t)})$ denotes the output response of agent $v_j^{m,(t)}$ at round t. Since an agent may be connected to multiple neighbors that contribute unequally, Eq. 5 normalizes the aggregation weights accordingly. This procedure applies analogously to temporal message passing. After aggregating spatio-temporal information, each node leverages the MLLM to update its response. After T rounds, the

summary agent synthesizes messages from the two modalities to produce the final answer.

Intra-Modal Edge Optimization. The edge optimization objective maximizes the expected utility while penalizing the nuclear norm of the softened adjacency matrices, thereby balancing task performance and communication-graph sparsity. Accordingly, the training objective is defined as follows:

$$\max_{\tilde{\mathbf{A}}_m^{\mathcal{S}}, \tilde{\mathbf{A}}_m^{\mathcal{T}}} \mathbb{E}_{\mathcal{G}^{\text{intra}}} \left[\phi(\mathcal{G}^{\text{intra}})\right] - \|\tilde{\mathbf{A}}_m^{\mathcal{S}}\|_* - \|\tilde{\mathbf{A}}_m^{\mathcal{T}}\|_* \tag{6}$$

subject to $\|\mathbf{A}_m^{\mathcal{S}} - \tilde{\mathbf{A}}_m^{\mathcal{S}}\|_F + \|\mathbf{A}_m^{\mathcal{T}} - \tilde{\mathbf{A}}_m^{\mathcal{T}}\|_F \le \delta$, where δ defines the noise tolerance level, $\phi(\cdot)$ denotes the task performance metric, and $\|\cdot\|_*$ denotes the nuclear norm. Because the optimization process may involve API calls to the underlying MLLM, which are non-differentiable, we adopt the policy gradient method [42] to approximate gradients. Following [48], we optimize the objective by sampling K graph instances from $\mathcal{G}^{\text{intra}}$, denoted by $\mathcal{G}_k^{\text{intra}}$. The sampling and optimization process is formulated as follows:

$$\nabla \mathbb{E}_{\mathcal{G}^{\text{intra}}}[\phi(\mathcal{G}^{\text{intra}})] \approx \frac{1}{K} \sum_{k=1}^{K} \phi(\mathcal{G}_k^{\text{intra}}) \cdot \nabla \log \mathcal{P}(\mathcal{G}_k^{\text{intra}}) \tag{7}$$

$$\mathcal{P}(\mathcal{G}_k^{\text{intra}}) = \left(\prod_{t=1}^{T} \prod_{e_{ij} \in \mathcal{E}_m^{\mathcal{S},k,(t)}} \tilde{\mathbf{A}}_m^{\mathcal{S},(t)}[i,j] \right) \cdot \left(\prod_{t=2}^{T} \prod_{e_{ij} \in \mathcal{E}_m^{\mathcal{T},k,(t)}} \tilde{\mathbf{A}}_m^{\mathcal{T},(t)}[i,j] \right) \tag{8}$$

where $\mathcal{P}(\mathcal{G}_k^{\text{intra}})$ is the sampling probability of $\mathcal{G}_k^{\text{intra}}$, and K is the number of sampled graphs.

4.2 Inter-Modal Graph Sparsification

Building on the intra-modal graphs, we integrate cross-modal information and learn the importance of cross-modal edges. The inter-modal topology is optimized for both task performance and structural regularity. Additionally, we introduce a modality alignment score to encourage consistent task understanding across modalities during sparsification.

Edge Weight Initialization. Recall that the intra-modal edges ($\mathcal{E}_m^{\mathcal{S}}, \mathcal{E}_m^{\mathcal{T}}$), adjacency matrices ($\mathbf{A}_m^{\mathcal{S}}, \mathbf{A}_m^{\mathcal{T}}$), and softened adjacency matrices ($\tilde{\mathbf{A}}_m^{\mathcal{S}}, \tilde{\mathbf{A}}_m^{\mathcal{T}}$) are inherited from the preceding intra-modal phase. Given predefined inter-modal spatial edges $\mathcal{E}_{m'}^{\mathcal{S}}$ and temporal edges $\mathcal{E}_{m'}^{\mathcal{T}}$, each agent is connected to all agents in the other modality, where $m' \in \{\text{txt} \to \text{vis}, \text{vis} \to \text{txt}\}$ denotes the two inter-modal directions. We obtain the initial inter-modal adjacency matrices $\mathbf{A}_{m'}^{\mathcal{S}}$ and $\mathbf{A}_{m'}^{\mathcal{T}}$. Similarly, we initialize the softened adjacency matrices of the inter-modal edges, $\tilde{\mathbf{A}}_{m'}^{\mathcal{S}}$ and $\tilde{\mathbf{A}}_{m'}^{\mathcal{T}}$, following Eq. 3.

Inter-Modal Message Passing. Each agent $v_i^{(t)}$ at round t receives neighboring information from both modalities through $\mathcal{G}^{\text{inter}}$. We first consider spatial message passing:

$$\mathcal{M}_{i,\text{inter}}^{\mathcal{S},(t)} = \sum_{v_j^{(t)} \in \mathcal{N}_{\text{inter}}^{\mathcal{S}}(v_i^{(t)})} \mathcal{W}^{(\mathcal{S},(t))}[i,j] \cdot \mathcal{O}(v_j^{(t)}), \tag{9}$$

$$\mathcal{W}^{(\mathcal{S},(t))}[i,j] = \frac{\exp(\tilde{\mathbf{A}}_{m'}^{\mathcal{S}}[i,j])}{\sum_{v_k^{(t)} \in \mathcal{N}_{\text{inter}}^{\mathcal{S}}(v_i^{(t)})} \exp(\tilde{\mathbf{A}}_{m'}^{\mathcal{S}}[i,k])}, \tag{10}$$

where $\mathcal{N}_{\text{inter}}^{\mathcal{S}}(v_i^{(t)})$ denotes the set of spatial neighbors of $v_i^{(t)}$ at round t. Aggregation weights are computed as in Eq. 10, analogous to Eq. 5. Temporal message passing is defined analogously. Each node aggregates spatio-temporal information from both modality directions. In addition, it incorporates the results of the intra-modal message passing discussed previously. Finally, the agent generates its output response based on the MLLM backbone. After T rounds of multi-agent discussion, the summary agent synthesizes messages from all agents to generate the final answer.

Inter-Modal Edge Optimization. Similar to intra-modal edge optimization, our training objective jointly optimizes task performance and graph sparsity. We additionally introduce a modality alignment score to encourage consistent semantic understanding across modalities by preserving inter-modal associations:

$$\max_{\tilde{\mathbf{A}}_{\text{inter}}} \mathbb{E}_{\mathcal{G}^{\text{inter}}}[\phi(\mathcal{G}^{\text{inter}})] - \sum_{\mathcal{X} \in \{\mathcal{S},\mathcal{T}\}} \sum_{m' \in \{\text{txt}\to\text{vis}, \text{vis}\to\text{txt}\}} \|\tilde{\mathbf{A}}_{m'}^{\mathcal{X}}\|_* + \sum_{\mathcal{X} \in \{\mathcal{S},\mathcal{T}\}} \mathcal{L}_{\text{align}}(\tilde{\mathbf{A}}_{\text{txt}\to\text{vis}}^{\mathcal{X}}, \tilde{\mathbf{A}}_{\text{vis}\to\text{txt}}^{\mathcal{X}}) \tag{11}$$

subject to $\sum_{\mathcal{X} \in \{\mathcal{S},\mathcal{T}\}} \|\mathbf{A}_{\text{txt}\to\text{vis}}^{\mathcal{X}} - \tilde{\mathbf{A}}_{\text{txt}\to\text{vis}}^{\mathcal{X}}\|_F \le \delta$ and $\sum_{\mathcal{X} \in \{\mathcal{S},\mathcal{T}\}} \|\mathbf{A}_{\text{vis}\to\text{txt}}^{\mathcal{X}} - \tilde{\mathbf{A}}_{\text{vis}\to\text{txt}}^{\mathcal{X}}\|_F \le \delta$, where δ is the noise tolerance level. $\tilde{\mathbf{A}}_{\text{inter}}$ denotes the collection of four softened adjacency matrices, namely $\tilde{\mathbf{A}}_{m'}^{\mathcal{S}}$ and $\tilde{\mathbf{A}}_{m'}^{\mathcal{T}}$ with $m' \in \{\text{txt} \to \text{vis}, \text{vis} \to \text{txt}\}$. The modality alignment score is defined as:

$$\mathcal{L}_{\text{align}}(\tilde{\mathbf{A}}_{\text{txt}\to\text{vis}}^{\mathcal{X}}, \tilde{\mathbf{A}}_{\text{vis}\to\text{txt}}^{\mathcal{X}}) = \frac{1}{N_T N_I} \sum_i \frac{\tilde{\mathbf{A}}_{\text{txt}\to\text{vis}}^{\mathcal{X}}[i,:] \cdot \left(\tilde{\mathbf{A}}_{\text{vis}\to\text{txt}}^{\mathcal{X}}\right)^\top [i,:]}{\|\tilde{\mathbf{A}}_{\text{txt}\to\text{vis}}^{\mathcal{X}}[i,:]\| \cdot \|\left(\tilde{\mathbf{A}}_{\text{vis}\to\text{txt}}^{\mathcal{X}}\right)^\top [i,:]\|}, \tag{12}$$

where N_T and N_I denote the numbers of textual and visual agents, respectively, and i indexes the rows of the corresponding matrices.

4.3 Progressive Edge Pruning

Our hierarchical pruning operates in two stages: it first sparsifies intra-modal edges within each modality and then prunes inter-modal connections between the textual and visual modalities. This yields a progressive pruning strategy in which the per-round pruning rate decays over rounds; however, because pruned edges are permanently removed, the communication graph becomes progressively sparser. Since pruning is determined during training, the communication topology used at inference time is the resulting pruned graph. For ease of exposition, we take the intra-modal spatial edges as an example. Recall that the adjacency matrix and its softened version are denoted as $\mathbf{A}_m^{\mathcal{S},(t)}$ and $\tilde{\mathbf{A}}_m^{\mathcal{S},(t)}$, respectively. The pruning masks $\mathbf{B}_m^{\mathcal{S},(t)}$ are defined as follows:

$$\mathbf{B}_m^{\mathcal{S},(t)} = \mathbb{1}\left(\mathbf{A}_m^{\mathcal{S},(t)} \neq 0 \wedge \text{Top}K\left(\tilde{\mathbf{A}}_m^{\mathcal{S},(t)}, \#(\mathbf{A}_m^{\mathcal{S},(t)}) \times (1 - p^{(t)})\right)\right), \tag{13}$$

where $\mathbb{1}(\cdot)$ is the indicator function. $\text{Top}K(\mathbf{X}, K)$ returns a Boolean matrix of the same shape as $\mathbf{X}$, whose entries indicate whether each element of $\mathbf{X}$ belongs to the top-K largest elements, with ties broken arbitrarily. $\#(X)$ denotes the number of elements in matrix X. The pruning rate at round t is defined as $p^{(t)} = p^{(t-1)} \cdot \exp\left(-\frac{t}{T}\right)$, following the setting in [48]. In the binary mask matrix $\mathbf{B}_m^{\mathcal{S},(t)}$, entries set to zero indicate that the corresponding adjacency-matrix elements are pruned, whereas entries set to one indicate preserved connections. Pruning is applied at the end of round t, and the resulting masks are used to update the adjacency matrices before

Table 1: Performance comparison between M^{3}Prune and baselines on the domain-specific ScienceQA benchmark.

Training Paradigm	Method	Llama3.2-VL (11B)									Qwen-VL-Max								
		Subject			Context Modality			Grade		Avg.	Subject			Context Modality			Grade		Avg.
		NAT	Soc	LAN	TXT	IMG	NO	G1-6	G7-12		NAT	Soc	LAN	TXT	IMG	NO	G1-6	G7-12	
Zero-shot	SP	82.92	88.79	77.55	82.87	82.47	79.09	84.30	79.99	$82.76_{\pm 1.2}$	91.90	90.56	89.36	91.47	88.09	90.99	91.44	90.10	$90.96_{\pm 0.9}$
	CoT	84.89	**97.95**	69.00	85.88	**93.70**	70.51	87.36	76.59	$83.51_{\pm 0.5}$	92.36	91.56	90.45	91.64	88.99	91.99	92.03	91.10	$91.70_{\pm 0.8}$
Single-agent RAG	Wiki-LLaVA	82.30	92.21	80.25	83.69	84.66	80.21	84.80	82.17	$83.85_{\pm 0.8}$	91.00	95.62	89.98	91.84	91.45	89.44	93.13	89.10	$91.69_{\pm 0.4}$
	RoRA-VLM	82.23	92.11	80.20	83.64	84.61	80.13	84.77	82.13	$83.80_{\pm 0.9}$	90.92	95.51	89.88	91.78	91.40	89.39	93.04	89.07	$91.65_{\pm 1.5}$
	EchoSight	82.35	92.23	80.27	83.70	84.69	80.23	84.81	82.21	$83.88_{\pm 1.1}$	91.04	95.65	90.00	91.87	91.47	89.48	93.19	89.12	$91.73_{\pm 0.8}$
	LLaVA-mR2AG	82.37	92.22	80.31	83.73	84.71	80.26	84.85	82.23	$83.90_{\pm 0.8}$	91.07	95.68	90.02	91.90	91.48	89.51	93.22	89.17	$91.80_{\pm 0.6}$
	CoRe-MMRAG	82.44	92.31	80.30	83.77	84.73	80.28	84.86	82.27	$83.94_{\pm 1.2}$	91.10	95.72	90.06	91.93	91.50	89.55	93.24	89.18	$91.82_{\pm 1.0}$
Multi-agent RAG	OmniSearch	82.71	88.13	**83.36**	82.61	80.56	83.33	85.25	81.80	$84.01_{\pm 0.3}$	92.50	93.43	90.09	93.35	91.31	89.71	93.35	89.76	$92.07_{\pm 0.4}$
	ViDoRAG	83.88	92.15	83.00	83.97	84.96	82.39	87.43	81.72	$85.39_{\pm 0.7}$	93.99	92.20	90.45	93.91	90.26	91.29	93.79	90.74	$92.70_{\pm 0.5}$
	HM-RAG	**86.25**	94.89	72.45	**86.56**	89.70	75.23	87.78	79.45	$84.80_{\pm 1.3}$	94.30	93.66	92.36	93.47	90.36	94.08	94.15	92.78	$93.66_{\pm 1.2}$
	E-Agent	83.00	87.58	83.09	82.45	80.40	83.51	85.31	81.59	$83.98_{\pm 1.4}$	93.67	93.44	89.18	94.06	91.84	89.90	93.20	91.13	$92.46_{\pm 1.5}$
	Ours	85.97	94.60	81.45	85.29	86.47	**83.83**	**88.25**	**83.65**	$\mathbf{86.61}_{\pm 0.6}$	**97.51**	**96.63**	**97.82**	**97.70**	**96.13**	**97.98**	**98.05**	**96.24**	$\mathbf{97.41}_{\pm 0.4}$

Table 2: Performance comparison on general-domain benchmarks.

Backbone	Training Paradigm	Method	Vidoseek												MultimodalQA				Average	
			Single-hop		Multi-hop		Text		Table		Chart		Layout		Image		Text			
			Acc*	EM	Acc*	EM	Acc*	EM	Acc*	EM	Acc*	EM	Acc*	EM	Acc*	EM	Acc*	EM	Acc*	EM
Llama3.2-VL (11B)	Zero-shot	SP	25.89	2.48	12.07	8.45	26.25	0.01	9.14	5.14	14.65	10.83	22.88	4.38	14.68	15.23	36.28	31.46	$20.23_{\pm 0.8}$	$9.75_{\pm 0.6}$
		CoT	30.61	5.17	7.44	9.24	21.00	4.25	10.14	5.86	9.92	13.83	25.25	6.01	16.82	15.68	33.42	32.53	$19.33_{\pm 1.3}$	$11.57_{\pm 1.1}$
	Single-agent RAG	Wiki-LLaVA	47.35	16.94	46.16	36.72	48.88	11.37	38.38	29.20	37.61	31.88	50.61	24.83	13.01	12.11	58.10	49.95	$42.51_{\pm 0.5}$	$26.63_{\pm 0.8}$
		RoRA-VLM	47.25	16.83	46.05	36.55	48.66	11.20	38.21	29.03	37.44	31.81	50.43	24.71	12.91	11.49	57.91	49.67	$42.37_{\pm 1.5}$	$26.41_{\pm 1.2}$
		EchoSight	47.29	16.90	46.08	36.62	48.75	11.25	38.29	29.14	37.58	31.85	50.55	24.79	12.95	11.59	57.97	49.75	$42.43_{\pm 0.5}$	$26.49_{\pm 0.4}$
		LLaVA-mR2AG	47.63	17.23	46.45	36.92	48.97	11.43	38.89	29.45	37.74	31.99	50.87	24.99	12.99	11.76	58.13	49.98	$42.71_{\pm 0.9}$	$26.72_{\pm 0.7}$
		CoRe-MMRAG	47.53	17.23	46.34	36.85	48.98	11.44	38.41	29.54	37.87	31.96	50.78	24.96	13.14	12.42	58.25	50.04	$42.66_{\pm 1.2}$	$26.81_{\pm 1.3}$
	Multi-agent RAG	OmniSearch	48.83	17.95	45.61	37.00	44.75	16.50	41.29	30.29	**40.94**	**38.39**	50.59	23.73	17.27	17.50	58.23	51.14	$43.44_{\pm 0.8}$	$29.06_{\pm 0.7}$
		ViDoRAG	49.37	**20.61**	45.55	36.75	56.00	**23.50**	41.57	27.00	38.20	32.02	50.32	27.30	20.23	18.64	64.30	56.33	$45.69_{\pm 1.0}$	$30.27_{\pm 1.1}$
		HM-RAG	**56.72**	20.22	34.33	32.52	**56.75**	19.50	37.00	30.29	34.02	33.66	**51.08**	23.37	**31.77**	24.09	53.84	49.11	$44.44_{\pm 1.5}$	$29.10_{\pm 0.9}$
		E-Agent	48.06	18.04	44.47	37.21	48.75	16.75	38.29	28.71	37.58	34.85	50.14	25.05	23.55	21.05	64.38	56.15	$44.40_{\pm 0.6}$	$29.73_{\pm 0.9}$
		Ours	44.65	19.38	**54.12**	**42.25**	51.25	12.50	**48.57**	**37.71**	40.13	34.39	50.41	**28.08**	29.55	**28.41**	**73.16**	**64.56**	$\mathbf{48.98}_{\pm 0.7}$	$\mathbf{33.41}_{\pm 0.8}$
Qwen-VL-Max	Zero-shot	SP	39.84	4.65	13.28	8.25	36.25	0.01	13.14	7.43	21.02	14.01	32.60	4.93	35.00	30.23	51.65	43.29	$30.35_{\pm 0.6}$	$14.10_{\pm 0.4}$
		CoT	45.89	5.41	15.29	11.05	43.75	3.25	12.57	6.57	26.11	16.65	37.53	6.79	32.27	27.73	53.92	44.94	$33.42_{\pm 0.9}$	$15.30_{\pm 0.6}$
	Single-agent RAG	Wiki-LLaVA	53.91	19.89	49.91	40.81	57.11	12.39	47.10	37.01	46.27	39.22	55.70	26.58	24.30	22.21	69.73	60.42	$50.50_{\pm 1.2}$	$32.32_{\pm 1.1}$
		RoRA-VLM	55.45	20.00	50.41	40.99	57.40	12.43	47.38	37.10	46.42	39.36	55.89	26.69	24.25	22.24	69.85	60.56	$50.88_{\pm 1.3}$	$32.42_{\pm 0.8}$
		EchoSight	55.81	20.16	50.50	41.05	57.50	12.50	47.43	37.14	46.50	39.49	56.03	26.99	24.32	22.27	69.87	60.63	$51.00_{\pm 0.6}$	$32.53_{\pm 0.5}$
		LLaVA-mR2AG	55.97	20.34	50.82	41.49	57.63	12.74	47.83	37.33	46.69	39.61	56.46	27.21	24.84	22.61	69.99	60.91	$51.28_{\pm 0.9}$	$32.78_{\pm 1.0}$
		CoRe-MMRAG	55.82	20.29	50.75	41.36	57.54	12.64	47.71	37.30	46.59	39.58	56.33	27.10	24.61	22.40	69.91	60.88	$51.16_{\pm 1.1}$	$32.69_{\pm 1.5}$
	Multi-agent RAG	OmniSearch	59.84	20.38	50.91	42.85	61.25	14.75	44.57	35.86	49.04	43.04	59.59	27.71	33.05	29.18	78.03	68.66	$54.54_{\pm 1.3}$	$35.30_{\pm 1.7}$
		ViDoRAG	68.43	22.29	58.73	48.67	84.25	29.00	50.43	40.57	60.96	53.04	66.01	28.52	30.68	30.45	82.41	74.30	$62.74_{\pm 0.8}$	$40.86_{\pm 1.1}$
		HM-RAG	66.88	20.71	52.50	46.03	68.00	20.50	49.29	39.43	47.59	44.94	65.33	28.27	36.09	33.45	78.27	69.27	$57.99_{\pm 1.0}$	$37.83_{\pm 0.9}$
		E-Agent	68.27	22.67	52.30	43.03	64.25	17.50	49.29	37.00	49.50	44.49	66.42	28.97	41.45	38.36	83.15	74.03	$59.33_{\pm 0.6}$	$38.26_{\pm 0.8}$
		Ours	**73.95**	**26.67**	**68.41**	**56.54**	**85.00**	**30.00**	**64.57**	**53.14**	**63.06**	**56.05**	**73.56**	**33.97**	**70.23**	**70.00**	**86.08**	**78.10**	$\mathbf{73.11}_{\pm 0.7}$	$\mathbf{50.56}_{\pm 0.9}$

Table 3: Dataset statistics.

Dataset	Training	Validation	Testing
MultimodalQA	23.8k	2.4k	3.66k
Vidoseek	–	–	1.1k
ScienceQA	12.7k	4.2k	4.2k

round $(t+1)$: $\mathbf{A}_m^{\mathcal{S},(t+1)} = \mathbf{A}_m^{\mathcal{S},(t)} \odot \mathbf{B}_m^{\mathcal{S},(t)}$ where $\odot$ denotes element-wise multiplication. The operations for the other adjacency matrices ($\mathbf{A}_m^{\mathcal{T},(t)}$, $\mathbf{A}_{m'}^{\mathcal{S}}$, and $\mathbf{A}_{m'}^{\mathcal{T}}$) are analogous. After T rounds, the multi-modal multi-agent communication flows are optimized for more efficient and effective inference.

5 Experiments

5.1 Experimental Settings

Datasets and Evaluation Metrics. In Table 3, we summarize three representative multi-modal datasets.[2] **MultimodalQA** [35] and **Vidoseek** [39] are general multi-modal question-answering datasets, while **ScienceQA** [26] is a multi-modal multiple-choice dataset tailored to scientific-domain questions. **MultimodalQA** contains multi-modal question–answer pairs involving tables, text, and images. In our experiments, we focus exclusively on the text and image modalities. Answering a question typically requires leveraging one or more relevant images and text passages selected from a pool of approximately 20 visual and textual distractors. **Vidoseek** is designed for visual document retrieval and question answering and serves as a large-scale benchmark for visual-document RAG systems. It addresses the limitations of single-image or single-document QA datasets and comprises roughly 6,000 images across 12 domains, including economics, technology, and literature. **ScienceQA** constitutes a multi-modal benchmark for scientific question answering, covering three core disciplines: Natural Sciences, Social Sciences, and Formal Sciences. Each sample combines a textual question with optional visual contexts such as diagrams, charts, or photographs.

For evaluation, we adopt both semantic-level and string-level metrics to assess answer quality. Specifically, **Acc*** measures semantic consistency between model responses and ground truths using DeepSeek-V3 [22] as the judge on a scale from 1 to 5; scores of 4 or higher are considered correct. **EM** denotes exact match,

[2]Since Vidoseek has no official training set, we use the training set of MultimodalQA due to the similarity of their multimodal QA distributions.

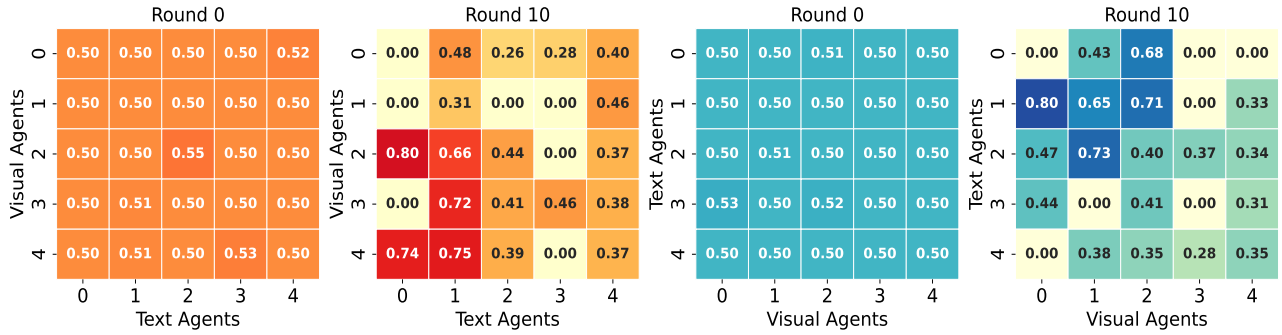


Figure 3: Evolution of communication edge weights from visual-to-text (Left) and text-to-visual (Right) agents on ScienceQA.

where a prediction is considered correct if it exactly matches or fully contains the ground truth.

Baselines. (1) Vanilla MLLM Prompting. We include Standard Prompting (SP) and Chain-of-Thought (CoT) prompting [41]. **MLLM-based Standard Prompting** generates an answer directly from the question. **MLLM-based Chain-of-Thought (CoT)** augments the instruction with intermediate reasoning steps, guiding the model to produce the final answer through step-by-step inference. **(2) Single-Agent RAG Methods. CoRe-MMRAG** [37] is an end-to-end four-stage framework that addresses knowledge inconsistencies in mRAG. **Wiki-LLaVA** [5] proposes a hierarchical three-stage RAG framework for MLLMs. **EchoSight** [44] employs a two-stage retrieval and re-ranking mechanism based on the similarity between multi-modal query tokens and text segments. **RORA-VLM** [32] uses a two-stage retrieval process: query images first retrieve visual entities, followed by textual query expansion for knowledge retrieval. It incorporates a noise-resistant generation mechanism with adversarial noise injection during training and query-oriented visual token filtering. **LLaVA-mR^{2}AG** [51] enhances MLLMs for knowledge-based VQA through retrieval reflection (determining whether knowledge is necessary) and relevance reflection (generating answers solely from relevant retrieved passages). **(3) Multi-Agent RAG Methods.** Existing multi-modal multi-agent RAG models typically use fixed communication topologies for agent collaboration. **OmniSearch** [19] dynamically decomposes complex multi-modal questions into sub-question chains with adaptive retrieval actions. **ViDoRAG** [39] uses a Gaussian-mixture-model-based hybrid retrieval mechanism to dynamically fuse textual and visual features and determine the retrieval quantity. **HM-RAG** [24] employs a hierarchical multi-agent architecture: (1) a Decomposition Agent breaks down complex queries, followed by parallel retrieval agents; and (2) a Decision Agent integrates multi-source evidence via consensus voting and expert-model refinement. **E-agent** [40] introduces a planning–execution multi-agent framework that fuses external knowledge with internal reasoning.

Implementation Details. For open-source backbone models, we use Qwen2.5-VL-7B [4] and Llama3.2-VL-11B [12]. For the closed-source backbone, we employ Qwen-VL-Max [43] via its official inference API. All experiments are conducted on a single NVIDIA A800 GPU. We set the number of communication rounds to $T = 2$, the number of graph samples to $K = 10$, the learning rate to $\eta = 0.1$, and the noise tolerance level to $\delta = 0.1$. We deploy 5 textual and 5 visual agents for ScienceQA, 4 textual and 4 visual agents for Vidoseek, and 5 textual agents with 4 visual agents for MultimodalQA. Based on the initial spatial adjacency matrix, we first perform DAG sampling on the graph to ensure sequential communication among agents. Model training proceeds in two stages: intra-modal graph training and inter-modal graph training, with each stage using 40 training instances sampled from the corresponding dataset.[3]

Table 4: Ablation study of M^{3}Prune.

Backbone	Method	MultimodalQA	ScienceQA	Average
Llama3.2-VL-11B	**M^{3}Prune**	53.14	86.61	69.88
	w/o $\mathcal{L}_{\text{align}}$	50.23	84.04	67.14
	w/o norm	52.23	85.10	68.67
	w/o $\mathcal{G}^{\text{inter}}$	50.44	82.61	66.53
	w/o $\mathcal{G}^{\text{intra}}$-txt	48.78	83.56	66.17
	w/o $\mathcal{G}^{\text{intra}}$-vis	51.81	80.24	66.03
Qwen-VL-Max	**M^{3}Prune**	76.40	97.41	86.91
	w/o $\mathcal{L}_{\text{align}}$	74.13	95.56	84.85
	w/o norm	75.47	96.05	85.76
	w/o $\mathcal{G}^{\text{inter}}$	73.85	94.27	84.06
	w/o $\mathcal{G}^{\text{intra}}$-txt	70.91	95.69	83.30
	w/o $\mathcal{G}^{\text{intra}}$-vis	74.83	92.06	83.45

5.2 Main Results

We evaluate M^{3}Prune against three categories of baselines on both general and domain-specific multi-modal QA tasks. The results are summarized in Table 1 and Table 2. We make the following observations: (1) SP and CoT prompting exhibit limited performance due to the absence of external knowledge integration. (2) Single-agent RAG methods consistently outperform zero-shot prompting, confirming the importance of leveraging retrieved contexts. (3) Multi-agent RAG frameworks generally surpass their single-agent counterparts, suggesting that inter-agent discussion facilitates the exchange of viewpoints and mitigates errors caused by imperfect single-agent retrieval. (4) M^{3}Prune outperforms fixed-topology multi-agent methods on both general and domain-specific tasks, indicating that intra-modal and inter-modal pruning reduce redundant communication and suppress noisy viewpoints that would otherwise disrupt collaboration. (5) The gains of M^{3}Prune are consistent across all three backbone models with different scales, ranging from 7B open-source models to closed-source MLLMs, supporting the robustness and generalizability of our framework across backbones.

[3]Due to the fast convergence of MLLM-based multi-agent system training, our experimental settings (e.g., the number of training instances) follow prior work [19, 39, 40].

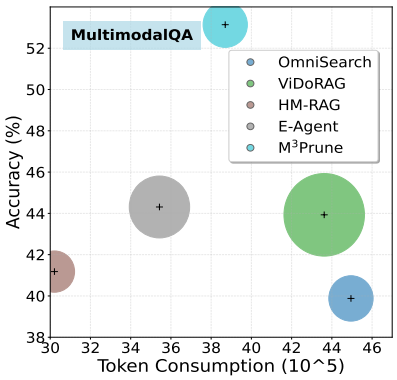


Figure 4: Token efficiency comparison (Acc./# Token) among multi-agent models. The total token count sums prompt and completion tokens.

5.3 Ablation Study

Table 4 presents an ablation study of the key components of M^{3}Prune. (1) ***w/o*** $\mathcal{L}_{align}$: Removing modality alignment leads to a clear performance drop, indicating that $\mathcal{L}_{\text{align}}$ helps align cross-modal semantics and stabilize inter-modal collaboration. (2) ***w/o norm***: Removing the norm regularization term weakens the model's control over overall edge density and reduces its ability to eliminate redundant edges. (3) ***w/o*** $\mathcal{G}^{inter}$: Disabling inter-modal graph sparsification reduces the method to aggregating intra-modal outputs without learned cross-modal communication. The resulting degradation suggests that structured inter-modal interaction is important for coherent multi-modal reasoning. (4) ***w/o*** $\mathcal{G}^{intra}$***-txt*** / ***w/o*** $\mathcal{G}^{intra}$***-vis***: We remove the contribution of the textual (respectively, visual) modality by replacing the outputs of the corresponding agents with dummy responses (e.g., "This is a dummy agent with no information."). The results show that both modalities contribute substantially, as removing either modality causes a significant drop in performance. Moreover, MultimodalQA is more sensitive to removing textual agents, whereas ScienceQA is more sensitive to removing visual agents, which is consistent with their task characteristics.

5.4 Detailed Analysis

(1) Communication Edge Evolution. To quantitatively analyze the evolution of the communication topology during training, we examine the inter-modal edge weights across discussion rounds. Specifically, we track the evolution of edge weights, i.e., the values of the softened adjacency matrices, from visual to textual agents (Fig. 3, left) and from textual to visual agents (Fig. 3, right) on ScienceQA using Qwen-VL-Max. As illustrated in Fig. 3, the learned differentiation of edge weights exhibits two properties: (1) The model progressively develops strong preferences for certain inter-agent connections as reasoning deepens, as reflected by the darker blocks. (2) Hierarchical graph pruning gradually sparsifies the communication topology by down-weighting low-correlation edges, as evidenced by the fading of non-essential connections. These observations suggest that the model learns to identify and reinforce semantically meaningful communication patterns while suppressing redundant interactions.

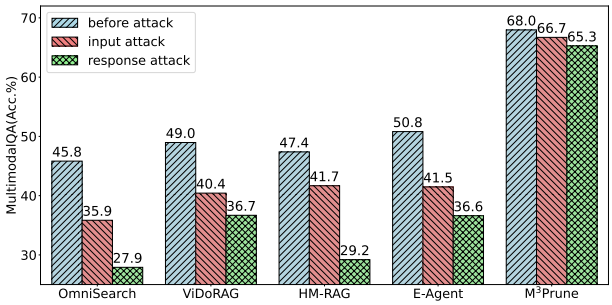


Figure 5: Model performance under adversarial attacks (prompt and response perturbations) on MultimodalQA.

(2) Token Efficiency Analysis. To evaluate the cost–performance trade-off of our framework, we analyze the relationship between accuracy and total token consumption using the token efficiency metric across all agents on MultimodalQA, Vidoseek, and ScienceQA with Llama3.2-VL-11B. As shown in Fig. 4, compared with advanced multi-modal multi-agent architectures, i.e., OmniSearch [19] and ViDoRAG [39], M^{3}Prune achieves higher accuracy with substantially fewer tokens. Relative to vanilla multi-agent baselines with simpler planning structures, e.g., E-agent [40] and HM-RAG [24], M^{3}Prune provides consistent accuracy gains with only a moderate increase in token usage. However, in terms of token efficiency (i.e., Acc./# Token), our model still outperforms these fixed-topology methods while using fewer tokens overall. This behavior stems from our adaptive hierarchical pruning, which selectively preserves informative communication clusters while removing redundant and noisy edges. In contrast, fixed-topology baselines may consume slightly fewer tokens in some cases but incur a clear drop in results.
(3) Robustness Verification. We evaluate whether M^{3}Prune maintains stable performance under adversarial conditions compared with fixed-topology baselines. We conduct experiments on MultimodalQA using Qwen2.5-VL-7B under two attack scenarios: (1) *Input Prompt Attack*: Replacing one agent with a malicious agent that ignores all inter-modal communications and relies solely on its own knowledge. (2) *Response Attack*: Substituting one agent with an adversarial agent that intentionally generates incorrect answers and misleading explanations to disrupt the reasoning of other agents. As shown in Fig. 5, fixed-topology multi-agent baselines suffer substantial performance degradation under both attacks, attributable to their inability to dynamically reweight interactions or isolate compromised agents. In contrast, M^{3}Prune maintains robust performance by adaptively pruning and reweighting edges to reduce the influence of compromised agents.

6 Conclusion

We present **M^{3}Prune**, a framework for multi-modal multi-agent systems that dynamically optimizes agent connections through structured pruning. Our approach addresses redundant communication in collaborative reasoning via progressive pruning over hierarchical intra-modal and inter-modal graphs. This two-stage strategy enables agents to first form task-specific perspectives within each modality, followed by semantically complementary interactions across modalities. Experiments on both general-domain and domain-specific benchmarks show that M^{3}Prune achieves state-of-the-art performance while substantially improving token efficiency.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62506110). It was also supported by the Natural Science Foundation of Anhui Province, China (Grant No. 2508085QF227) and the Hefei University of Technology Scientific Research Innovation Start-up Special Project Type A (Grant No. JZ2025HGQA0137).

References

[1] Pravallika Abbineni, Saoud Aldowaish, Colin Liechty, Soroosh Noorzad, Ali Ghazizadeh, and Morteza Fayazi. 2025. MuaLLM: A Multimodal Large Language Model Agent for Circuit Design Assistance with Hybrid Contextual Retrieval-Augmented Generation. *CoRR* abs/2508.08137 (2025).

[2] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics*. 16776–16809.

[3] Arian Askari, Emmanouil Stergiadis, Ilya Gusev, and Moran Beladev. 2025. HotelMatch-LLM: Joint Multi-Task Training of Small and Large Language Models for Efficient Multimodal Hotel Retrieval. In *Proceedings of the Association for Computational Linguistics*. 607–619.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv:2502.13923* (2025).

[5] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1818–1826.

[6] Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. 2023. ObjectFusion: Multi-modal 3D Object Detection with Object-Centric Fusion. In *IEEE/CVF International Conference on Computer Vision*. 18021–18030.

[7] Guillem Capellera, Antonio Rubio, Luis Ferraz, and Antonio Agudo. 2025. Unified Uncertainty-Aware Diffusion for Multi-Agent Trajectory Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22476–22486.

[8] Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. MAIN-RAG: Multi-Agent Filtering Retrieval-Augmented Generation. In *Proceedings of the Association for Computational Linguistics*. 2607–2622.

[9] Jiawei Chen, Yue Jiang, Dingkang Yang, Mingcheng Li, Jinjie Wei, Ziyun Qian, and Lihua Zhang. 2024. Can LLMs' Tuning Methods Work in Medical Multimodal Domain?. In *Medical Image Computing and Computer Assisted Intervention*. 112–122.

[10] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8415–8424.

[11] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2025. Progressive Multimodal Reasoning via Active Retrieval. In *Proceedings of the Association for Computational Linguistics*. 3579–3602.

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv:2407.21783* (2024).

[13] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *IEEE/CVF International Conference on Computer Vision*. 20349–20360.

[14] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv:1611.01144* (2016).

[15] Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J. Su, Camillo Jose Taylor, and Tanwi Mallick. 2025. Towards Rationality in Language and Multimodal Agents: A Survey. In *Proceedings of the Nations of the Americas Chapter of the Association for Computational Linguistics*. 3656–3675.

[16] Aisha Urooj Khan, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels da Vitoria Lobo, and Mubarak Shah. 2021. Found a Reason for me? Weakly-supervised Grounded Visual Question Answering using Capsules. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8465–8474.

[17] Jungsoo Lee, Janghoon Cho, Hyojin Park, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. 2025. Generalized Contrastive Learning for Universal Multimodal Retrieval. *CoRR* abs/2509.25638 (2025).

[18] Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2025. HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases. In *Proceedings of the Association for Computational Linguistics*. 879–893.

[19] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. 2025. Benchmarking Multimodal Retrieval Augmented Generation with Dynamic VQA Dataset and Self-adaptive Planning Agent. In *International Conference on Learning Representations*.

[20] Junxiong Lin, Yan Wang, Zeng Tao, Boyang Wang, Qing Zhao, Haorang Wang, Xuan Tong, Xinji Mai, Yuxuan Lin, Wei Song, Jiawen Yu, Shaoqi Yan, and Wenqiang Zhang. 2024. Adaptive Multi-modal Fusion of Spatially Variant Kernel Refinement with Diffusion Model for Blind Image Super-Resolution. In *European Conference on Computer Vision*. 363–380.

[21] Zudi Lin, Erhan Bas, Kunwar Yashraj Singh, Gurumurthy Swaminathan, and Rahul Bhotika. 2023. Relaxing Contrastiveness in Multimodal Representation Learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. 2226–2235. doi:10.1109/WACV56688.2023.00226

[22] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv:2412.19437* (2024).

[23] Hongjun Liu, Yinghao Zhu, Yuhui Wang, Yitao Long, Zeyu Lai, Lequan Yu, and Chen Zhao. 2025. MedMMV: A Controllable Multimodal Multi-Agent Framework for Reliable and Verifiable Clinical Reasoning. *CoRR* abs/2509.24314 (2025).

[24] Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. *CoRR* abs/2504.12330 (2025).

[25] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems*.

[26] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* (2022), 2507–2521.

[27] Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. LexLIP: Lexicon-Bottlenecked Language-Image Pre-Training for Large-Scale Image-Text Sparse Retrieval. In *IEEE/CVF International Conference on Computer Vision*. 11172–11183.

[28] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2025. DrVideo: Document Retrieval Based Long Video Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18936–18946.

[29] Tianyi Men, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Agent-RewardBench: Towards a Unified Benchmark for Reward Modeling across Perception, Planning, and Safety in Real-World Multimodal Agents. In *Proceedings of the Association for Computational Linguistics*. 17521–17541.

[30] Jingwei Peng, Jiehao Chen, Mateo Alejandro Rojas, and Meilin Zhang. 2025. MV-CoRe: Multimodal Visual-Conceptual Reasoning for Complex Visual Question Answering. *CoRR* abs/2508.07023 (2025).

[31] Akhil Perincherry, Jacob Krantz, and Stefan Lee. 2025. Do Visual Imaginations Improve Vision-and-Language Navigation Agents?. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3846–3855.

[32] Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. RoRA-VLM: Robust Retrieval-Augmented Vision Language Models. *CoRR* abs/2410.08876 (2024).

[33] Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. 2023. Decouple Before Interact: Multi-Modal Prompt Learning for Continual Visual Question Answering. In *IEEE/CVF International Conference on Computer Vision*. 2941–2950.

[34] Leonardo Ranaldi, Federico Ranaldi, and Giulia Pucci. 2025. R2-MultiOmnia: Leading Multilingual Multimodal Reasoning via Self-Training. In *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[35] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv:2104.06039* (2021).

[36] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24827–24837.

[37] Yang Tian, Fan Liu, Jingyuan Zhang, Victoria W., Yupeng Hu, and Liqiang Nie. 2025. CoRe-MMRAG: Cross-Source Knowledge Reconciliation for Multimodal RAG. In *Proceedings of the Association for Computational Linguistics*. 32967–32982.

[38] Guohua Wang, Shengping Song, Wuchun He, and Yongsen Zheng. 2025. CMHKF: Cross-Modality Heterogeneous Knowledge Fusion for Weakly Supervised Video Anomaly Detection. In *Proceedings of the Association for Computational Linguistics*. 31594–31607.

[39] Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. ViDoRAG: Visual Document Retrieval-Augmented Generation via Dynamic Iterative Reasoning Agents. *CoRR* abs/2502.18017 (2025).

[40] Yuechen Wang, Yuming Qiao, Dan Meng, Jun Yang, Haonan Lu, Zhenyu Yang, and Xudong Zhang. 2025. Efficient Agent: Optimizing Planning Capability for Multimodal Retrieval Augmented Generation. *CoRR* abs/2508.08816 (2025).

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[42] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[43] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765* (2025).

[44] Yibin Yan and Weidi Xie. 2024. EchoSight: Advancing Visual-Language Models with Wiki Knowledge. In *Findings of Empirical Methods in Natural Language Processing*. 1538–1551.

[45] Shuo Yang, Caren Han, Siwen Luo, and Eduard H. Hovy. 2025. MAGIC-VQA: Multimodal And Grounded Inference with Commonsense Knowledge for Visual Question Answering. In *Findings of the Association for Computational Linguistics*. 16967–16986.

[46] Xinlei Yu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Ruolin Shen, Jiangning Zhang, Xiaobin Hu, Yanwei Fu, and Shuicheng Yan. 2025. Visual Document Understanding and Question Answering: A Multi-Agent Collaboration Framework with Test-Time Scaling. *CoRR* abs/2508.03404 (2025).

[47] Zhengrong Yue, Shaobin Zhuang, Kunchang Li, Yanbo Ding, and Yali Wang. 2025. V-Stylist: Video Stylization via Collaboration and Reflection of MLLM Agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3195–3205.

[48] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2025. Cut the Crap: An Economical Communication Pipeline for LLM-based Multi-Agent Systems. In *International Conference on Learning Representations*.

[49] Shiyue Zhang, Zheng Chong, Xujie Zhang, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. 2024. GarmentAligner: Text-to-Garment Generation via Retrieval-Augmented Multi-level Corrections. In *European Conference on Computer Vision (Lecture Notes in Computer Science, Vol. 15083)*. 148–164.

[50] Taolin Zhang, Dongyang Li, Qizhou Chen, Chengyu Wang, and Xiaofeng He. 2025. BELLE: A Bi-Level Multi-Agent Reasoning Framework for Multi-Hop Question Answering. In *Proceedings of the Association for Computational Linguistics*. 4184–4202.

[51] Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. 2024. mR^{2}AG: Multimodal Retrieval-Reflection-Augmented Generation for Knowledge-Based VQA. *CoRR* abs/2411.15041 (2024).

[52] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025. MultiAgentBench : Evaluating the Collaboration and Competition of LLM agents. In *Proceedings of the Association for Computational Linguistics*. 8580–8622.

[53] Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models. *CoRR* abs/2410.03577 (2024).