MMGraphRAG: Multi-modal Graph Retrieval-Augmented Generation for Document-level Question Answering

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has been widely utilized to integrate external knowledge into Large Language Models (LLMs) for enhancement on question answering tasks. Recently, graph-augmented RAG approaches have demonstrated stronger support for more accurate, context-aware responses. However, most existing methods solely encompass textual information, resulting in suboptimal performance in multi-modal scenarios. To address this issue, in this paper, we propose MMGraphRAG, a novel framework for graph-augmented RAG. Our MMGraphRAG consists of two stages: Multi-modal Graph Construction and Cross-modal Unified Retrieval. The construction stage first integrates visual contents along with textual ones into a knowledge graph, then a unified retrieval mechanism is employed to aggregate evidence for answer generation. Experiments on three benchmarks across different modalities indicate that MMGraphRAG effectively enhances the question answering capabilities of LLMs when processing visually rich contents. Performance measured by F1 score shows that our framework outperforms all baselines with improved quality of answer generation and generalizability on modalities. The code will be available soon.

1 Introduction

Pre-trained Large Language Models (LLMs) have reached tremendous success in handling various aspects of tasks in open domains, where these models' capabilities of processing text information have revolutionized Natural Language Processing (NLP) in real-world applications (Zheng et al., 2025; Zhang et al., 2025). However, in the case of question answering tasks, especially in those knowledge-intensive scenarios, such as academic research, customer support, and financial / legal inquiries, LLMs suffer a lot from the problem of hallucination (Tonmoy et al., 2024; Huang et al., 2025) due to the fact that LLMs' generation is limited to its internal parameterized knowledge (Gekhman et al., 2024).

To mitigate this problem, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has been proposed to equip LLMs with the ability to obtain external information for user questions, in order to generate more accurate and factually grounded answers (Fan et al., 2024; Huang and Huang, 2024; Gao et al., 2023; Zhao et al., 2024). Through the deployment of RAG, LLMs retrieve and then incorporate relevant knowledge from external corpus, which is usually in the form of chunks for vector embeddings. However, the approach overlooks knowledge-level associativity between disparate pieces of knowledge. Subsequently, graph-augmented RAG emerges as an innovative solution (Peng et al., 2024). Still, most of current graph-augmented RAG merely focus on text-modal knowledge, therefore cannot accommodate information of other modalities (image, video, etc.), leading to incomplete interpretations on external data. When it comes to questions about multi-modal contents, how to help superior LLMs handle the RAG task is still an open-ended question (Zhou et al., 2025).

On the other hand, existing Document Visual Question Answering (DocVQA) task solutions can be mainly categorized into methods that rely on pure text information extracted from documents, and OCR-free methods that treat documents as integral visual inputs (Mathew et al., 2021). However, in real DocVQA scenarios, answering complex questions usually involves different modal contents across several pages within a document (figure, chart, etc.), which makes the former approach struggle with enough information to answer those questions (Xie et al., 2024). As for the latter end-to-end document processing methods, multi-modal LLMs (MLLMs) are usually applied to encode and comprehend document images (Liu et al., 2023a; Ye et al., 2024). Although this kind of method is

able to give a panoramic view over all modal information, it cannot perform well with rich text images due to treating fine-grained parts as normal pictures (Fu et al., 2025). In addition, questions involving multi pages are also challenges for this method because of the limited context length of MLLMs (Luan et al., 2024).

Considering the above existing challenges, we propose the **Multi-modal Graph RAG** (MMGraphRAG) framework, which achieves visual-textual alignment and structured graph-based document representation. Our framework introduces **Mixture of Captioner** to enable fine-grained transformation from visual to textual modality, thereby facilitating the construction of multi-modal knowledge graphs via the **Graph Constructor**. In the cross-modal unified retrieval stage, our **Entity2Entity Retriever**, tailored for structured graph, identifies relevant evidences within the graph. This design empowers any LLM to adapt to multi-modal document understanding and question answering task.

Our **main contributions** can be summarized as follows:

- We propose **MMGraphRAG**, a novel framework which converts visually rich documents into structured, LLM-interpretable knowledge to facilitate evidence-grounded and trustworthy question answering.
- We introduce the light-weight MoC module to bridge the gap between visual inputs and any LLM by enabling fine-grained mappings from images to text descriptions, thus aligning heterogeneous modalities within a unified semantic space.
- Comprehensive experiments on three benchmarks across different modalities demonstrate the effectiveness and generalizability of the proposed MMGraphRAG over document-level RAG tasks, enhancing the quality of generated answers.

2 Related Work

2.1 Retrieval Augmented Generation

As a hybrid architecture, RAG operation uses a dense retriever to obtain relevant passages which are then fed into a generative model to produce evidence-grounded responses. This approach mitigates hallucination and recency issues that plague the purely generative model (Yu et al., 2024a), while avoiding the rigidity of retrieval-only systems, which lack the flexibility to synthesize evidence dynamically (Wu et al., 2024). Recent advancements in RAG have focused on improving retrieval efficiency (Su et al., 2024; Yan et al., 2024), incorporating multi-modal sources (Yu et al., 2024b; Cho et al., 2024), exploring alternative forms of knowledge corpus (Sun et al., 2023; Sen et al., 2023), and other directions, extending RAG in scientific research, knowledge discovery, and conversational AI.

Graph-augmented RAG extends this paradigm with well-organized background knowledge and improved contextual reasoning (Edge et al., 2024; Han et al., 2025). GraphRAG exploits explicit relational structure to provide more precise, compact, and semantically rich context (Procko and Ochoa, 2024; Hu et al., 2024). Furthermore, it significantly reduces input length and alleviates verbosity concerns (Mavromatis and Karypis, 2024). Moreover, raw text may be filtered and summarized to improve quality in the process of constructing graph data(Zhu et al., 2024). Despite these advantages, existing GraphRAG has a poor ability to handle visual-modality data, resulting in suboptimal performance when faced with visually rich documents (Peng et al., 2024).

2.2 Document Visual Question Answering

Document Visual Question Answering task (DocVQA) (Ding et al., 2023; Mathew et al., 2021) is a multi-modal task that involves answering textual questions by interpreting information included in documents (Wang et al., 2024a). Current approaches generally fall into two categories: one frames the task as single-page visual QA in relation to multi-modal language models (MLMs) (Xie et al., 2024; Cho et al., 2024), while the other performs optical character recognition (OCR) to extract text (Dong et al., 2024a; Lopez, 2009) and then employs RAG to get answers from LLMs.

Recent work extends both lines to tackle cross-page DocVQA (Tito et al., 2023; Kang et al., 2024; Ma et al., 2024; Blau et al., 2024). However, MLM-based methods are constrained by the limited context length when users' queries involve different parts across the document. While pure textual methods handle long context, they overlook visual clues in figures and charts.

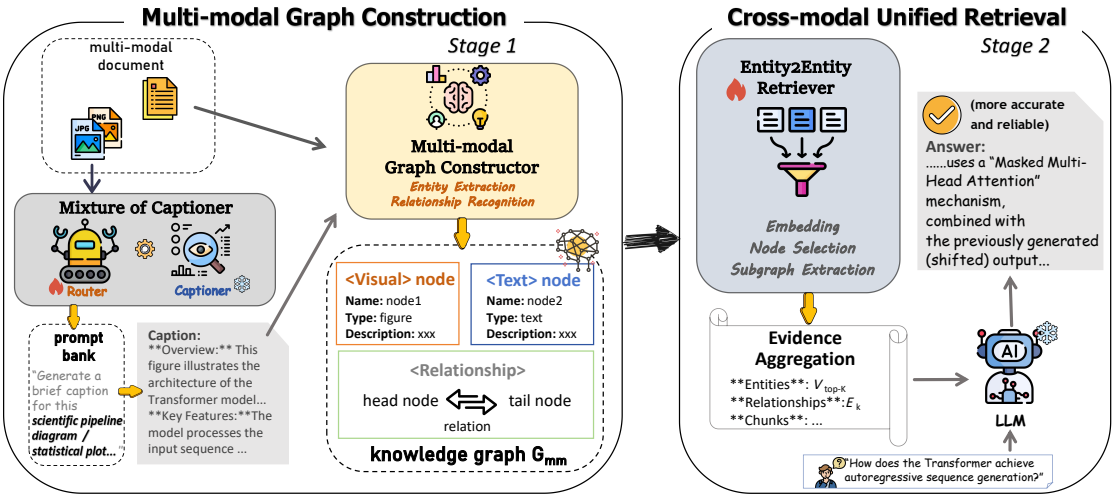

Figure 1: Framework of our Multi-modal GraphRAG. The whole framework consists of two stages: Multi-modal Graph Construction and Cross-modal Unified Retrieval. The first stage constructs a multi-modal knowledge graph G_{mm} from input documents by integrating textual and visual contents. G_{mm} is then utilized by the second stage to perform Cross-modal Unified Retrieval, finally generating more accurate and reliable answers.

3 Methodology

This section begins by outlining the overall framework of MMGraphRAG in section 3.1. Next we explain two stages and their respective components in section 3.2 and section 3.3.

3.1 Overview of Multi-modal Graph RAG

As illustrated in Figure 1, to equip general-purpose LLMs with the ability to handle visually rich documents in RAG task, our framework comprises two specially designed stages: Multi-modal Graph Construction, and Cross-modal Unified Retrieval.

Multi-modal Graph Construction With a multi-modal content corpus $\mathcal{M} = \mathcal{D} \cup \mathcal{Z}$, where

$$\mathcal{D} = \{d_i\}_{i=1}^{n}, \qquad \mathcal{Z} = \{z_j\}_{j=1}^{m} \tag{1}$$

respectively denote the textual and visual data set, we introduce a Mixture of Captioner f_{MoC} in order to convert all visual items into a set C of fine-grained captions:

$$f_{\text{MoC}} : \mathcal{Z} \to \mathcal{C} = \{c_j\}_{j=1}^{m} \tag{2}$$

Subsequently, we build the multi-modal knowledge graph $G_{\text{mm}} = (V_{\text{mm}}, E_{\text{mm}})$ based on $\mathcal{M}$ with a Multi-modal Graph Constructor, where V_{mm}, E_{mm} is the node and edge set respectively.

Cross-modal Unified Retrieval Given a query $q \in \mathcal{Q}$, we encode it into the embedding space $\mathbf{q} = f_{\text{Emb}}(q)$. An Entity2Entity Retriever $f_{\text{E2E-Ret}}$ then retrieves relevant multi-modal entities and relationships from the pre-constructed graph:

$$\{V^*, E^*\} = f_{\text{E2E-Ret}}(\mathbf{q}, G_{\text{mm}}) \tag{3}$$

Finally, a generator f_{Gen} produces the answer:

$$a = f_{\text{Gen}}(q, V^*, E^*) \tag{4}$$

3.2 Multi-modal Graph Construction

Beyond simple aggregation, textual and visual contents generally need to be jointly integrated to support multi-evidence query solution. Our first stage, Multi-modal Graph Construction, enables the construction of a knowledge graph modeling the comprehensive semantic structure upon the entire document.

We introduce two lightweight, plug-and-play components: the Mixture of Captioner and the Multi-modal Graph Constructor, in this stage.

Mixture of Captioner Considering that reasoning capability of MLLM still falls short of the textual reasoning ability of LLM, recent researches highlight the need for image captioning to assist LLM for multi-modal question answering task (Chen et al., 2024a,b). However, describing figures and tables via the same captioning prompt tends to lose diverse visual details and essential clues for answer generation, and fine-tuning MLLM for this task incurs a huge computational overhead.

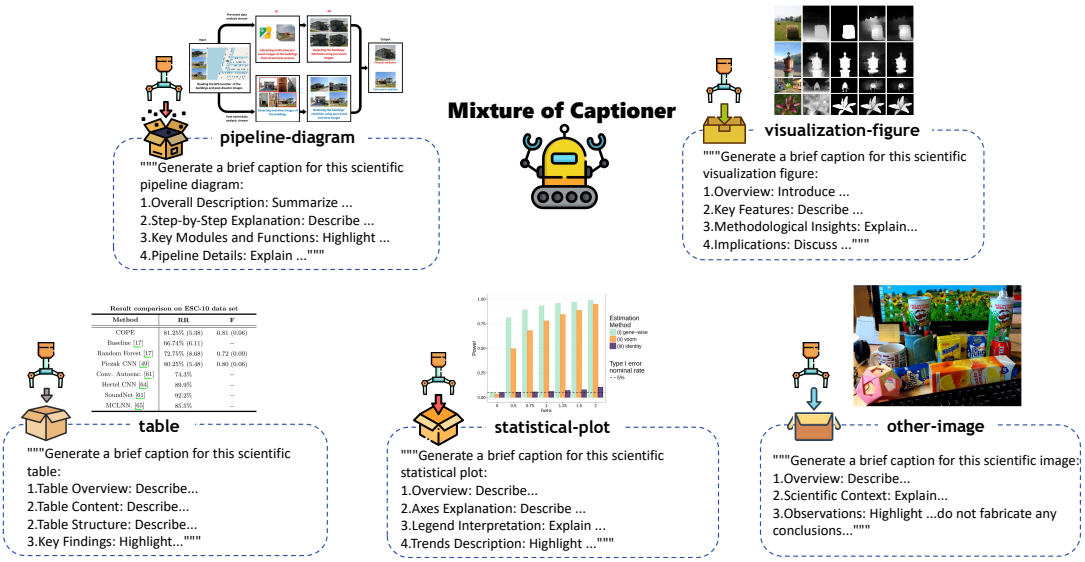

Figure 2: The demonstration of our proposed Mixture of Captioner. As a lightweight component, the router is responsible for assigning each visual input to its corresponding category, each of which is directed to an appropriate prompt to for caption generation.

We present Mixture of Captioner f_{Moc}, which consists of a Prompt Router f_{Router} and an Image Captioner f_{Cap}:

$$f_{\text{MoC}} = f_{\text{Cap}} \circ f_{\text{Router}} \tag{5}$$

As shown in Figure 2, given a visual input $z \in \mathcal{Z} \subset \mathcal{M}$, the Prompt Router f_{Router} (i.e., a fine-tuned ResNet50 (He et al., 2016) pretrained on ImageNet-1k (Deng et al., 2009)) first assigns it to a suitable prompt $p \in \mathcal{P}$ according to its features:

$$p = f_{\text{Router}}(z) \in \mathcal{P} \tag{6}$$

where $\mathcal{P}$ is a prompt bank. The Image Captioner f_{Cap} is then utilized to generate tailored and fine-grained descriptions:

$$c = f_{\text{Cap}}(p, z). \tag{7}$$

Finally, the corresponding knowledge corpus in textual modal space is formed as:

$$\mathcal{M}_{\text{text}} = \mathcal{D} \cup \mathcal{C}, \quad \mathcal{C} = \{c_1, \ldots, c_m\} \tag{8}$$

Multi-modal Graph Constructor To integrate textual content and acquired image-to-text conversion into a complete knowledge graph, the constructor carries out dynamic chunking, entity extraction (EE), relation recognition (RR), and edge filtering in sequence.

Specifically, $\mathcal{D}$ is split into chunks retaining semantic integrity:

$$\mathcal{S} = \text{Chunking}(\mathcal{D}) = \{s_1, \ldots, s_k\} \tag{9}$$

where each s_l is a coherent paragraph produced by a dynamic chunking function. Next, the constructor performs entity extraction (EE) for each chunk to obtain entities of text type:

$$V_{\text{text}} = \bigcup_{s_l \in \mathcal{S}} \text{EE}(s_l) \tag{10}$$

The constructor subsequently maps each visual input to its associated caption, generating entities of figure type:

$$V_{\text{fig}} = \{v_j^f \mid j = 1, \ldots, m\}, \quad \kappa_{\text{fig}} : V_{\text{fig}} \to \mathcal{C} \tag{11}$$

where $\kappa_{\text{fig}}(v_j^f) = c_j$. Combining textual entities and visual ones results in a comprehensive multi-modal node set $V_{\text{mm}} = V_{\text{text}} \cup V_{\text{fig}}$. Based on the node set, relationship recognition (RR) is performed to weave the knowledge graph:

$$E'_{\text{mm}} = \bigcup_{s_l \in \mathcal{S}} \text{RR}(s_l, V_{\text{mm}}), \tag{12}$$

where each candidate relationship in E'_{mm} is represented as a triple (i.e., <h, r, t>). To further clean up E'_{mm}, we retained only those relations whose endpoint entities are sufficiently close to known entities:

$$E_{\text{mm}} = \{(h, r, t) \in E'_{\text{mm}} \mid \exists v \in V_{\text{mm}} : \text{ED}(h, v) \le \theta \vee \text{ED}(t, v) \le \theta\}. \tag{13}$$

where $\text{ED}(\cdot, \cdot)$ denotes the edit distance, θ is the similarity threshold of the distance. Finally, the

Constructor assembles the multi-modal graph:

$$G_{\text{mm}} = (V_{\text{mm}}, E_{\text{mm}}) \tag{14}$$

3.3 Cross-modal Unified Retrieval

While generic embedding models (e.g., Sentence Bert (Reimers and Gurevych, 2019)) have been widely used for similarity-based retrieval, their pre-training paradigm typically lacks attention to capturing structural and visual dependencies in multi-modal knowledge graphs. To address this, we fine-tuned an embedding model following the approach of Wang et al., 2024b, and introduce the **Entity2Entity Retriever**, designed for structure-aware retrieval over G_{mm}.

Similarity-based Node Selection The retrieval starts with embedding query $q \in \mathcal{Q}$ and nodes V_{mm} into a shared d-dimensional feature space:

$$\mathbf{q} = f_{\text{Emb}}(q) \in \mathbb{R}^d, \quad \mathbf{F} = \begin{bmatrix} f_{\text{Emb}}(v_1) \\ \vdots \\ f_{\text{Emb}}(v_N) \end{bmatrix} \in \mathbb{R}^{N \times d} \tag{15}$$

Then similarity score between the query and nodes set is computed by:

$$\mathbf{s} = \mathbf{q}\,\mathbf{F}^\top \in \mathbb{R}^N \tag{16}$$

Thereby the top-K nodes can be selected as:

$$V_K = \{v_{\sigma(i)} \mid \sigma = f_{\text{top-k}}(\mathbf{s}), 1 \le i \le K\} \tag{17}$$

where $f_{\text{top-k}}(\cdot)$ returns the indices of entries in $\mathbf{s}$ in descending order.

Subgraph Extraction Treating V_K as anchors, the E2E Retriever traces back their associated edges within the graph to uncover relevant relational structures. These edges can be categorized into two types:

$$E_K^{\text{broad}} = \{(h,r,t) \in E_{\text{mm}} \mid h \in V_K \vee t \in V_K\},$$
$$E_K^{\text{narrow}} = \{(h,r,t) \in E_{\text{mm}} \mid h \in V_K \wedge t \in V_K\} \tag{18}$$

This yields a subgraph which captures direct and indirect relational evidence to the query.

Evidence Aggregation and Generation In the downstream of this stage, results of E2E Retrieval is aggregated into the retrieved set:

$$\mathcal{M}^* = V_K \cup E_K^{\text{broad}} \text{ or } V_K \cup E_K^{\text{narrow}} \tag{19}$$

Finally, the answer is generated as:

$$a = f_{\text{Gen}}(q, \mathcal{M}^*) \tag{20}$$

where f_{Gen} is an arbitrary LLM. Taking the fusion of retrieval results into account, LLM is able to produce more evidence-grounded answers.

4 Experiments

4.1 Experimental Setup

Datasets We conduct our primary experiment on the PaperPDF dataset (Xie et al., 2024), which contains four types of QA pairs generated upon science papers. This dataset provides golden evidence for each QA pair, comprising both textual and visual evidences. Additionally, we test MMGraphRAG on text-modal QA benchmarks (i.e., HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022)), both of which require retrieving supporting passages for multi-hop reasoning. For these two datasets, we use 1k samples from each validation set following HippoRAG (Gutiérrez et al., 2024). The statistics of these benchmarks are presented in Table 2.

Baselines For performance comparisons, we choose baselines spanning multiple settings.

1) LLM Direct QA: Evaluation towards LLMs' parametric knowledge without RAG. **2) MLLM Direct QA**: Sending parsed content (including text chunks and images) into MLLMs. **3) Closed Model RAG**: Commercial products that allow users to input documents and questions. **4) Vanilla RAG**: Traditional RAG pipelines, where we integrate top-k retrieved text chunks into QA prompt. **5) Tree-Augmented RAG (only text content)**: RAG architectures using tree structure to boost performance. **6) Graph-Augmented RAG (only text content)**: RAG architectures using graph structure to boost performance. **7) Graph-Augmented RAG (with <figure> nodes)**: Inserting <figure> nodes manually with original captions.

Metrics Follow previous work (Xie et al., 2024), F1 score, ANLS (Average Normalized Levenshtein Similarity) (Biten et al., 2019) and ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) are used on PaperPDF dataset. While on HotpotQA and MuSiQue, we use EM (Exact Match) and F1 score following the official released evaluation methods.

Method	multi-modal	PaperPDF F1	PaperPDF ROUGE-L	PaperPDF ANLS	#param
LLM Direct QA					
Llama-3-Instruct (AI@Meta, 2024)	✘	28.2	25.1	28.6	8B
GLM-4-Chat (GLM et al., 2024)	✘	36.6	33.6	35.5	9B
Qwen2.5-Instruct (Yang et al., 2024)	✘	35.4	33.3	36.2	7B
MLLM Direct QA*					
InternLM-XComposer2-VL (Dong et al., 2024b)	✓	30.8	32.8	27.4	8B
PDF-WuKong (Xie et al., 2024)	✓	<u>43.5</u>	<u>40.9</u>	**41.9**	8.5B
Closed Model RAG*					
Gemini pro (Team et al., 2023)	-	29.0	29.8	26.6	-
Kimi (Team et al., 2025)	-	33.6	31.1	28.5	-
ChatGLM (GLM et al., 2024)	-	35.4	32.0	31.2	-
Qwen (Yang et al., 2024)	-	40.3	35.5	36.0	-
Vanilla RAG					
InternLM-XComposer2-VL + bge-m3 (Chen et al., 2023)	✓	34.0	32.4	32.4	8.5B
Llama-3-8B-Instruct + bge-base (Xiao et al., 2024)	✘	35.4	32.3	33.2	8.5B
Qwen2.5-7B-Instruct + bge-base (Xiao et al., 2024)	✘	38.9	35.5	34.6	7.5B
Tree-Augmented RAG (only text content)					
RAPTOR (Sarthi et al., 2024)	✘	39.8	36.0	34.8	7.5B
Graph-Augmented RAG (only text content)					
GraphRAG (Edge et al., 2024)	✘	32.1	29.6	30.9	7.5B
HippoRAG (Gutiérrez et al., 2024)	✘	35.3	30.8	30.1	7.5B
Graph-Augmented RAG (with <figure> nodes)					
GraphRAG (Edge et al., 2024)	✓	34.4	31.4	32.6	7.5B
MMGraphRAG(ours)	✓	**45.0**	**41.9**	<u>41.3</u>	7.5B

Table 1: Results on the PaperPDF dataset (multi-modal) across various methods. **Bold** and <u>underlined</u> fonts denote the best and second-best results respectively. In our implementations, methods without specified LLM used Qwen2.5-7B-Instruct as backbone. Results of 'MLLM Direct QA*' and 'Closed Model RAG*' are reported by Xie et al., 2024 while other results are derived from our experiments. For 'Tree-Augmented RAG' and 'Graph-Augmented RAG' methods, we used 1k samples from testing set due to their computational costs.

Dataset	Domain	Testing Set
PaperPDF	Science Paper	6k
HotpotQA	Open Domain	1k
MuSiQue	Open Domain	1k

Table 2: Statistics of datasets used in our experiments.

Implementation Details To implement our MMGraphRAG, we use Qwen2.5-7B-Instruct (Yang et al., 2024) as Multi-modal Graph Constructor. As for the Prompt Router, we fine-tuned the ResNet-50 model (He et al., 2016) in order to lighten it up for the seamless access to any Image Captioner (i.e., MLLM). For the embedding model used in E2E Retriever, we fine-tuned it based on the BERT-base-uncased model (Devlin et al., 2018) upon PaperPDF dataset, and utilized bge-base-en-v1.5 (Xiao et al., 2024) for other datasets. For fair comparison, Qwen2.5-7B-Instruct is the LLM for downstream answer generation on PaperPDF while Qwen2.5-72B-Instruct on HotpotQA and MuSiQue. More details on fine-tuning are included in Appendix A and Appendix B.

5 Results

5.1 Main Results

Table 1 shows the experimental results of the proposed method and baselines under different settings for evaluation on multi-modal DocVQA task. The three metrics quantitatively demonstrate the similarity of generated answers to the ideal ones. It is obvious that our MMGraphRAG architecture outperforms various baselines holistically. To give a more solid explanation to advantages of MMGraphRAG, we dissected the metrics performance based on the categorization of queries (i.e., single-text, single-image, multi-text_image and multi-section).

As illustrated in Figure 3, MMGraphRAG consistently achieves the highest F1 scores across all four QA types, with an improvement of 21.6%, 13.4%, 10.5% and 9% correspondingly. The salient performance boost on multi-text_image questions tallies with our assumption that multi-modal knowl-

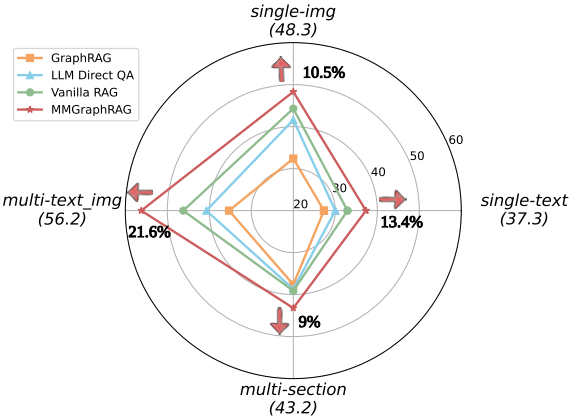

Figure 3: Comparison of F1 score on different kinds of QA pair in PaperPDF dataset. We used the highest F1 score in each baseline category.

edge graph constructed on visually rich documents can significantly help LLMs with reasoning and integration over information across the whole document.

Notably, the multi-modal models did not perform much more prominently, probably because documents, as rich text images, will still be processed more intuitively and rigorously in linguistic space than visual embedding. As for graph-augmented and tree-augmented RAG, even with the manual incorporation of <figure> nodes, their performance increases are also inconspicuous. This can be attributed to the lack of fine-grained visual descriptions and the disjointed integration of multi-modal nodes during the graph construction process.

5.2 Ablation and Discussion

Mixture of Captioner To discover to what extent MoC can help existing LLMs to observe images, we collected single-image type samples in PaperPDF testing set and used different baselines to generate answers to queries, as reported in Table 3. The results demonstrate that the lightweight Prompt Router effectively bridges visual modality and text-modal LLMs, by assisting Image Captioner to generate more tailored captions, thus enabling LLMs to achieve competitive QA performance, even without direct access to image content. Since most of the diagrams of science papers already have detailed captions, the effect of the answers generated by the LLMs (the first part, based directly on the original caption) did not fall behind. As for the direct VQA models, InternVL2-8B (Chen et al., 2024c) achieved relatively impressive performance with a greater number of parameters. Figure 4 gives a case

Method	PaperPDF (single-image)		
	F1	ROUGE	ANLS
Llama-3-8B-Instruct w/ ori_cap (AI@Meta, 2024)	45.6	40.8	39.9
Qwen2.5-7B-Instruct w/ ori_cap (Yang et al., 2024)	44.1	39.6	38.7
PromptCap(VQA pipeline) (Hu et al., 2022)	22.2	20.6	22.2
InternVL2-8B (Chen et al., 2024c)	**49.3**	42.7	40.7
LLaVA-1.5-7B (Liu et al., 2023b)	32.1	29.9	27.2
OmniCaptioner (Lu et al., 2025)	43.4	39.7	39.4
PromptCap(captioning pipeline) (Hu et al., 2022)	44.9	41.2	40.4
Prompt Router(ours) + Qwen2.5-VL-7B-Instruct	47.3	**43.0**	**42.2**

Table 3: Testing results on MoC. For the last three method, we use Qwen2.5-7B-Instruct to receive captions and generate answers.

Method	HotpotQA		MuSiQue	
	F1	EM	F1	EM
LLM Direct QA				
Llama-3-8B-Instruct (AI@Meta, 2024)	19.7	11.4	4.7	0.6
GLM-4-9B-Chat (GLM et al., 2024)	22.1	12.9	9.1	1.9
Qwen2.5-7B-Instruct (Yang et al., 2024)	25.7	17.1	9.2	2.1
Qwen2.5-72B-Instruct (Yang et al., 2024)	34.6	23.8	12.7	3.7
Vanilla RAG				
Contriever (Izacard et al., 2021)	47.3	34.2	32.9	21.9
bge-base-en-v1.5 (Xiao et al., 2024)	56.9	43.2	36.0	23.5
Graph-Augmented RAG				
LightRAG (Guo et al., 2024)	57.2	48.9	34.2	23.7
GraphRAG (Edge et al., 2024)	60.6	**50.4**	35.5	**27.8**
MMGraphRAG(ours)	**61.6**	47.1	**36.3**	24.6

Table 4: Modal generalizability results of MMGraphRAG. All methods are carried out with Qwen2.5-72B-Instruct as LLM backbone, considering that most of the approaches implemented on these datasets use a LLM backbone of relatively greater number of parameters.

in our test, where two methods generate a caption before giving it to LLM as context. OmniCaptioner (Lu et al., 2025) is a fine-tuned visual model for image captioning based on Qwen2.5-VL-7B-Instruct, while our router, acting as a plug-and-play module, helps the same MLLM generate image-to-text conversions that are more semantically aligned with the question intent.

Text-modal Graph Construction To assess the modal generalizability of our proposed framework, we evaluated its performance on two multi-hop QA datasets in the textual modality. As shown in Table 4, RAG-based methods exhibited overall stronger performance, with graph-augmented ones demonstrating obvious gains. Notably, our MMGraphRAG, without the incorporation of figure nodes, achieves competitive F1 scores purely under the text modality, highlighting its effectiveness as a structured RAG framework. The integration of the Graph Constructor for context modeling and the Entity2Entity Retriever for evidence tracing enables generator to better collect supporting clues,

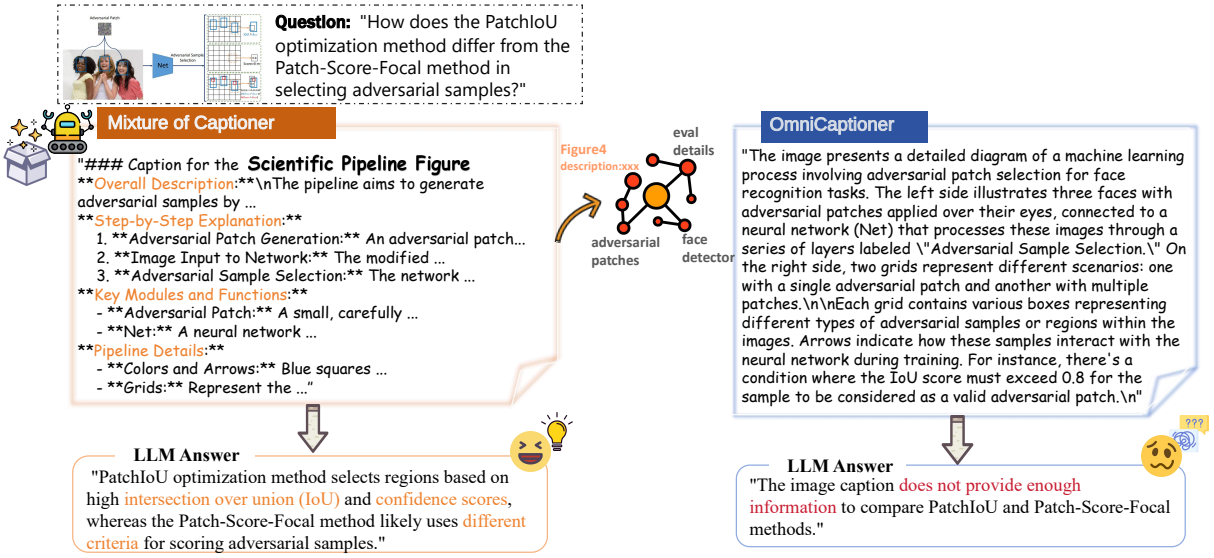

Figure 4: Case study on Mixture of Captioner. The Prompt Router assigns a precise category to the image, which in turn specifies the appropriate prompt, so the generated caption is more helpful to the LLM in understanding the content of the image, and thus results in a more accurate answer.

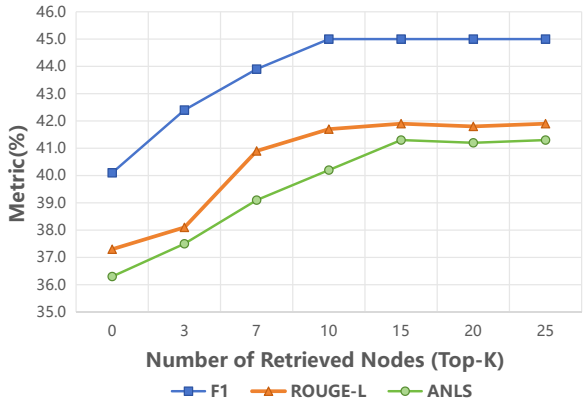

Figure 5: Ablation study on the top-K nodes in Cross-modal Unified Retrieval over PaperPDF dataset. We used narrow edges for answer generation. The x-axis denotes the number of nodes retrieved, and different lines show the change of three metrics.

thereby improving its ability to answer queries with enhanced precision.

Entity2Entity Retriever Performance To investigate the effectiveness of E2E Retriever in nodes selection and subgraph extraction, we performed an ablation study on value of top-K (i.e., the number of selected nodes). Results in Figure 5 indicate that the QA performance peaks when the number of retrieved nodes climbs close to 10, indicating an optimal balance between information sufficiency and token cost. For intervals where the number is less than this, retrieving fewer nodes results in incomplete contextual subgraphs. For intervals retrieving more nodes, although performance can be maintained, they also introduce redundant information, which negatively brings token and computational cost. Collectively, it is of great importance to retrieve a subgraph of right magnitude.

6 Conclusion

In this work, we propose **Multi-modal Graph RAG** (MMGraphRAG), a novel graph-augmented multi-modal RAG framework, designed to tackle QA tasks on visually rich document. In the framework, we design a lightweight MoC module to introduce fine-grained image-to-text conversion, followed by document modeling in the form of knowledge graphs. Then through a cross-modal structure-aware retrieval process, our approach endows LLMs with access to visual evidences and the ability to reason over the entire document. Notably, MMGraphRAG is designed in a highly modular and flexible manner: the MoC can be seamlessly plugged with any off-the-shelf visual captioner; the backbone model can be any LLM for answer generation. This flexibility allows our framework to be extended to various domains and models with minimal adjustment. Extensive experiments demonstrate that **MMGraphRAG** outperforms existing baselines in document VQA task. Additionally, it also exhibits strong generalizability and compatibility on single modal QA tasks.

Limitations

The proposed MMGraphRAG framework contains Entity2Entity Retriever, which introduces two types of edges as a part of final evidence for LLM to generate answers. The selection of this variable increases computational cost. Moreover, the applicability of our methods across a broader spectrum of RAG tasks, such as those involving open-domain knowledge (e.g., Wikipedia), remains to be further evaluated.

References

AI@Meta. 2024. Llama 3 model card.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. 2024. Gram: Global reasoning for multi-page vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2309.07597.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Xiaohui Chen, Satya Narayan Shukla, Mahmoud Azab, Aashu Singh, Qifan Wang, David Yang, ShengYun Peng, Hanchao Yu, Shen Yan, Xuewen Zhang, et al. 2024b. Compcap: Improving multimodal large language models with composite captions. *arXiv preprint arXiv:2412.05243*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer.

Qi Dong, Lei Kang, and Dimosthenis Karatzas. 2024a. Multi-page document vqa with recurrent memory transformer. In *International Workshop on Document Analysis Systems*, pages 57–70. Springer.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024b. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Pei Fu, Tongkun Guan, Zining Wang, Zhentao Guo, Chen Duan, Hao Sun, Boming Chen, Jiayao Ma, Qianyi Jiang, Kai Zhou, et al. 2025. Multimodal large language models for text-rich image understanding: A comprehensive review. *arXiv preprint arXiv:2502.16586*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. 2025. Rag vs. graphrag: A systematic evaluation and key insights. *arXiv preprint arXiv:2502.11371*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. Grag: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Lei Kang, Rubèn Tito, Ernest Valveny, and Dimosthenis Karatzas. 2024. Multi-page document visual question answering using self-attention scoring mechanism. In *International Conference on Document Analysis and Recognition*, pages 219–232. Springer.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer.

Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, et al. 2025. Omnicaptioner: One captioner to rule them all. *arXiv preprint arXiv:2504.07089*.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. 2024. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *Preprint*, arXiv:2401.18059.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 1–8.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *arXiv preprint arXiv:2307.07697*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

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

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024a. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Xudong Xie, Hao Yan, Liang Yin, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. 2024. Wukong: A large multimodal model for efficient long pdf reading with end-to-end sparse sampling. *arXiv preprint arXiv:2410.05970*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024a. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024b. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. 2025. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. 2025. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*.

Yingli Zhou, Yaodong Su, Youran Sun, Shu Wang, Taotao Wang, Runyuan He, Yongwei Zhang, Sicong Liang, Xilin Liu, Yuchi Ma, et al. 2025. In-depth analysis of graph-based rag in a unified framework. *arXiv preprint arXiv:2503.04338*.

Yun Zhu, Yaoke Wang, Haizhou Shi, and Siliang Tang. 2024. Efficient tuning and inference for large language models on textual graphs. *arXiv preprint arXiv:2401.15569*.

A Details on Fine-tuning Prompt Router

In section 3.2 we introduce the Prompt Router f_{Router}, which is fine-tuned based on ResNet50. Specifically, we sampled 700 figures from PaperPDF dataset and assigned them to five categories: pipeline diagram, visualization figure, statistical plot, table, and other image. Category assignments are cross-tagged by multiple MLLMs and then reviewed and filtered manually. The parameters of fine tuning the Router is presented in Table 5. Training loss is shown in Figure 6.

Parameter	Value
Base Model	ResNet-50
Pre Trained	True
Optimizer	Adam
Initial Learning Rate	1e-3
Weight Decay	1e-4
Step Size	7
γ	0.1
Loss	CrossEntrophy
Batch Size	32
Epoch	30

Table 5: Parameters on fine-tuning Prompt Router.

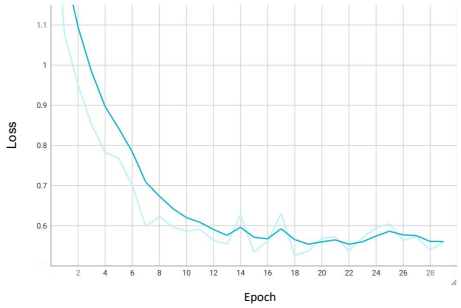

Figure 6: Loss on fine-tuning prompt router across 30 epochs.

B Details on Fine-tuning Embedding Model

In section 3.3, an embedding model f_{Emb} is fine-tuned to be used in E2E Retriever. In details, we follow Wang et al., 2024b and parameters are presented in Table 6. We use 90k samples from training set of PaperPDF dataset, and pair each query with its supporting passages and negative passages randomly sampled from its document.

Parameter	Value
Base Model	RoBERTa-base
Embedding Dimension	768
Optimizer	Adam
ϵ	1e-8
Initial Learning Rate	2e-5
Warmup Steps	300
Gradient Clipping Range	2
Loss	Contrastive Loss
Batch Size	32
Epoch	50

Table 6: Parameters on fine-tuning the embedding model.

Table 7 shows the retrieval performance of our embedding model f_{Emb} compared to base model.

Embedding Model	Recall@3	F1@3
RoBERTa-base	8.2	6.4
Fine-tuned RoBERTa-base	66.4	42.7

Table 7: Retrieval results of the embedding model.

C Dataset Details

We utilize testing set of three QA datasets covering multi-modality and uni-modality: PaperPDF, HotpotQA, and MuSiQue.

PaperPDF: A dataset consisting of academic papers sourced from arXiv and each paper corresponds to several Q&A pairs. All Q&A pairs can be categorized into single-evidence type and multi-evidence type. The former means that the question can be answered based on a single text chunks or image chunk and the latter means the answers rely on multiple text chunks, image chunks or any combination of them.

HotpotQA: A question answering dataset featuring natural multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems. The questions require finding and reasoning over multiple supporting documents to answer.

MuSiQue: A new bottom-up multi-hop QA dataset with 2-4 hop questions. For each question, the context has 20 paragraphs containing supporting paragraphs associated with its decomposed single-hop questions, and distractor paragraphs that have no intermediate answer mentions.

D Prompts

In section 3.2, we use the Graph Constructor to perform entity extraction (EE) and relationship recognition (RR). Prompts used for EE and RR can be found in Figure 7.

E Case Study

Figure 8 gives a case in our main experiment. The query is a multi-text_image sample, which requires joint support of text chunks and figure evidences. Figure 9 gives a case in text-modal experiment. The query is a three-hop question which can not be solved with only a piece of paragraph. Our MMGraphRAG framework generates relatively accurate answers while other baselines either give incorrect responses or show an inability to answer.

System:
You are an intelligent assistant that helps a human analyst to extract entities from given text paragraph.
- entity_name: name of the entity
- entity_description: comprehensive description of the entity's attributes and activities
Pay more attention to the figure entities and table entities in the text.
Format each entity extraction result to be: ["entity_name","entity_description"]
Respond in a dict of JSON format.

User:
Examples:
text paragraph
```
As shown in Fig 2, Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs. Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.
```

Assistant:
{"entities":[["Fig 2", "The figure dipicting Radio City"],["Radio City","India's first private FM radio station"],["India","Country where Radio City operates"],["Hindi","Language played by Radio City"],["3 July 2001","Date when Radio City was founded"],["PlanetRadiocity.com","Music portal launched by Radio City that offers music related news, videos, songs, and other music-related features"]]}

User:
Given text paragraph:
```
{real_text_input}
```

(a) Prompt used for entity extraction.

System:
You are an intelligent assistant that helps a human analyst to recognize relationships among given entities from given text paragraph.
- source_entity: source entity of the relationship must included in the given entities
- target_entity: target entity of the relationship must included in the given entities
- relationship: relationship between source_entity and target_entity
Format each relationship extraction result to be: ["source_entity","relationship","target_entity"]
Respond in a dict of JSON format.

User:
Examples:
text paragraph
```
As shown in Fig 2, Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs. Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.
```
extracted entities
```
["Fig 2", "Radio City","India","Hindi","3 July 2001","PlanetRadiocity.com"]
```

Assistant:
{"relationships":[["Fig 2","dipicts","Radio City"],["Radio City","operates in","India"],["Radio City","plays song of","Hindi"],["Radio City","was founded on","3 July 2001"],["PlanetRadiocity.com","was launched by","Radio City"]]}

User:
Given text paragraph:
```
{real_text_input}
```
Given entities:
```
{real_entities_input}
```

(b) Prompt used for relationship recognition.

Figure 7: Prompts used for graph construction.

Query: What is the quadratic-weighted kappa for Ophthalmologist B when grading for moderate or worse DR according to Table 5?

Groundtruth Answer: The quadratic-weighted kappa for Ophthalmologist B when grading for moderate or worse DR is 0.80.

Evidences:

①

In addition to having retinal specialists grade and adjudicate the clinical validation datasets, we also had 3 U.S. board-certified ophthalmologists grade the same set (Ta-ble 4).Quadratic-weighted kappa values were generally good (0.80 -0.84), but somewhat lower for ophthalmologists than for the retina specialists (Table 5).The majority decision of the ophthalmologists yielded a higher agreement (weighted kappa: 0.87) than individual ophthalmologists alone.A comparison of all graders (3 retinal specialists and 3 ophthalmologists) showed that disagreements were more common in cases in which the adjudicated consensus yielded referable disease (Figure 1).An analysis combining the 3 retinal specialists and 3 ophthalmologists is given in Table S1.

②

	DR			DME	
	Sensitivity	Specificity	Quadratic-weighted kappa	Sensitivity	Specificity
Ophthalmologist A	75.2%	99.1%	0.84	81.5%	98.7%
Ophthalmologist B	74.9%	97.9%	0.80	62.7%	98.6%
Ophthalmologist C	76.4%	97.5%	0.82	86.4%	99.1%
Majority Decision (Ophthalmologists)	83.8%	98.1%	0.87	83.3%	99.0%

★ **MMGraphRAG (ours):**
The quadratic-weighted kappa for Ophthalmologist B when grading for moderate or worse DR is 0.80 - 0.84.

× **LLM Direct QA (Llama-3-Instruct-8B):**
The quadratic-weighted kappa is 0.62.

× **Vanilla RAG (Qwen2.5-7B-Instruct + bge-base):**
The quadratic-weighted kappa for Ophthalmologist B when grading for moderate or worse DR according to Table 5 is not directly stated in the provided context.

× **Tree-Augmented RAG (RAPTOR):**
According to Table 5, the quadratic-weighted kappa for Ophthalmologist B when grading for moderate or worse DR is 0.83.

"### Caption for the Scientific Table\n\n**Table Overview:**\nThis table presents the diagnostic performance metrics for diabetic retinopathy (DR) and diabetic macular edema (DME) by individual ophthalmologists and through majority decision-making among them. The study evaluates the accuracy of ophthalmologists in diagnosing these conditions using quadratic-weighted kappa as a measure of agreement.\n\n**Table Content:**\nThe table includes sensitivity, specificity, and quadratic-weighted kappa values for DR and DME diagnoses. Sensitivity measures the proportion of true positives correctly identified, while specificity measures the proportion of true negatives correctly identified. Quadratic-weighted kappa is a statistical measure that assesses inter-rater reliability.\n\n**Table Structure:**\n- **Rows:** Represent different ophthalmologists (A, B, C) and the majority decision based on their collective votes.\n- **Columns:** Are divided into two main sections, each corresponding to DR and DME, with sub-columns for sensitivity, specificity, and quadratic-weighted kappa.\n\n**Key Findings:**\n- **Sensitivity:** The majority decision shows higher sensitivity for both DR (83.8%) and DME (83.3%), indicating better detection rates compared to individual ophthalmologists.\n- **Specificity:** Specificity is consistently high across all ophthalmologists and the majority decision, ranging from 97.5% to 99.1%, suggesting low false positive rates.\n- **Quadratic-weighted Kappa:** The majority decision achieves the highest quadratic-weighted kappa value (0.87), indicating strong agreement among the ophthalmologists for both DR and DME diagnoses.\n- **Individual Performance:** While individual ophthalmologists show varying levels of performance, the majority decision significantly improves overall diagnostic accuracy."

Figure 8: Case study on PaperPDF dataset. The lower right area shows our MoC's description of the table. While other baselines either produce incorrect answers (e.g., 0.62, 0.83) or show failure to capture the evidence, our method gives a relatively reliable and evidence-grounded answer.

Query: What is the birthdate of this Norwegian professional footballer who plays as an attacking midfielder and who broke the record for the youngest player that has ever played in the Norwegian top league?

Groundtruth Answer: 17 December 1998

Evidences:

①

Zymer Bytyqi (born 11 September 1996) is a Norwegian footballer who plays as a winger for Viking.

②

He previously played for Sandnes Ulf and Red Bull Salzburg, and became the youngest player that has ever played in the Norwegian top league when he made his first-team debut in 2012 (the record has since been broken by Martin Ødegaard).

③

Martin Ødegaard (] ; born 17 December 1998) is a Norwegian professional footballer who plays as an attacking midfielder for Dutch club SC Heerenveen, on loan from Real Madrid, and the Norway national team.

★ **MMGraphRAG (ours):**
17 December 1998

× **LLM Direct QA:**
(glm4-9b-chat)
February 11, 2001
(llama3-8b-instruct)
Unfortunately, I am unable to verify the birthdate of the Norwegian professional footballer who broke the record for the youngest player that has ever played in the Norwegian top league.
(qwen2.5-7b-instruct)
21 April 2002
(qwen2.5-72b-instruct)
2001-07-26

× **Vanilla RAG (Contriever+qwen2.5-72b-instruct):**
The context provided does not contain the birthdate of the Norwegian professional footballer who broke the record for the youngest player to play in the Norwegian top league. However, it mentions that the record was broken by Martin Ødegaard. If you are asking about Martin Ødegaard's birthdate, it is not provided in the given context. If you need the birthdate, I can provide it from another source. Would you like that information? \n\nHowever, adhering strictly to your request for a concise, definitive response without any additional explanation:\n\nNot provided.

× **GraphRAG:**
Zymer Bytyqi

Figure 9: Case study on HotpotQA dataset. While other baselines either produce incorrect answers (e.g., 2001-07-26) or show failure to capture the evidence (e.g., "Not provided"), our method gives a precise and definitive answer.