

MMGraphRAG: Bridging Vision and Language with Interpretable Multimodal Knowledge Graphs

Xueyao Wan, Hang Yu

1185909349@qq.com, hang023@e.ntu.edu.sg

Abstract

Retrieval-Augmented Generation (RAG) enhances language model generation by retrieving relevant information from external knowledge bases. However, conventional RAG methods face the issue of missing multimodal information. Multimodal RAG methods address this by fusing images and text through mapping them into a shared embedding space, but they fail to capture the structure of knowledge and logical chains between modalities. Moreover, they also require large-scale training for specific tasks, resulting in limited generalizing ability. To address these limitations, we propose MMGraphRAG, which refines visual content through scene graphs and constructs a multimodal knowledge graph (MMKG) in conjunction with text-based KG. It employs spectral clustering to achieve cross-modal entity linking and retrieves context along reasoning paths to guide the generative process. Experimental results show that MMGraphRAG achieves state-of-the-art performance on the DocBench and MMLongBench datasets, demonstrating strong domain adaptability and clear reasoning paths.

Introduction

Large Language Models (LLMs) have advanced in natural language generation, yet hallucination—factual inconsistency—remains a major limitation (Roller et al. 2020; Liu et al. 2024a; Huang et al. 2025). Due to their static parametric nature, LLMs cannot promptly integrate specialized and up-to-date data, causing incomplete or inaccurate knowledge in professional fields like medicine and law (Chu et al. 2025; Chen et al. 2025; Xia et al. 2025; Hindi et al. 2025). Retrieval-Augmented generation (RAG) mitigates this by incorporating external knowledge bases, retrieving relevant documents to enrich the generative process with up-to-date context, thereby reducing hallucinations (Lewis et al. 2020a). However, real-world information often co-exists in multimodal forms such as text, images, and tables. Text-only RAG methods cannot fully exploit visual information, leading to incomplete retrieval results (Lin et al. 2023).

Multimodal RAG (MRAG) has emerged to address this by typically mapping images and text into a shared embedding space to conduct cross-modal retrieval based on se-

mantic similarity. However, this approach struggles to capture the structure of knowledge and logical chains across modalities (Yu et al. 2024; Fayssse et al. 2025; Cho et al. 2024; Ling et al. 2025). Moreover, due to the inherent heterogeneity between different modalities, MRAG approaches require large amounts of domain-specific training to extract and align modality-specific features, which hinders generalizability. This specialization, along with noisy evidence in a shared embedding space, can mislead the model into confidently generating incorrect answers. (Mei et al. 2025). For instance, M3DocRAG achieves only 5.8% accuracy on questions with “unanswerable” ground truths, underscoring significant weaknesses in factual consistency (Cho et al. 2024). Thus, the limitations of MRAG in structured reasoning, generalizability, and hallucination reduction call for further research.

To bridge the gap, we propose MMGraphRAG, the first MRAG framework based on knowledge graph (KG). Specifically, we first refine image information into entities and relations through scene graphs. Then, we combine these scene graphs with text-based KG through cross-modal fusion to construct a fine-grained multimodal knowledge graph (MMKG). A crucial component of cross-modal fusion is Cross-Modal Entity Linking (CMEL), which enables the establishment of connections between entities from different modalities. However, research in this area remains in its early stage, and there is a lack of benchmarks to evaluate the task comprehensively. To address this, we build and release the CMEL dataset. Motivated by the challenges of accurate candidate selection in this task, we employ a spectral clustering algorithm to generate candidate entities for every visual entities, effectively enhancing the accuracy of the CMEL task. Finally, context is retrieved along multimodal reasoning paths within the MMKG to guide the generative process. Experimental results demonstrate that MMGraphRAG significantly outperforms existing RAG baselines on the DocBench (Zou et al. 2024) and MMLongBench datasets (Ma et al. 2025b), achieving strong domain adaptability without any training, while providing interpretable reasoning paths for multimodal inference.

Our main contributions are as follows:

1. **First multimodal GraphRAG framework:** We propose MMGraphRAG, which refines images into scene graphs and combines them with text-based KG to build a unified

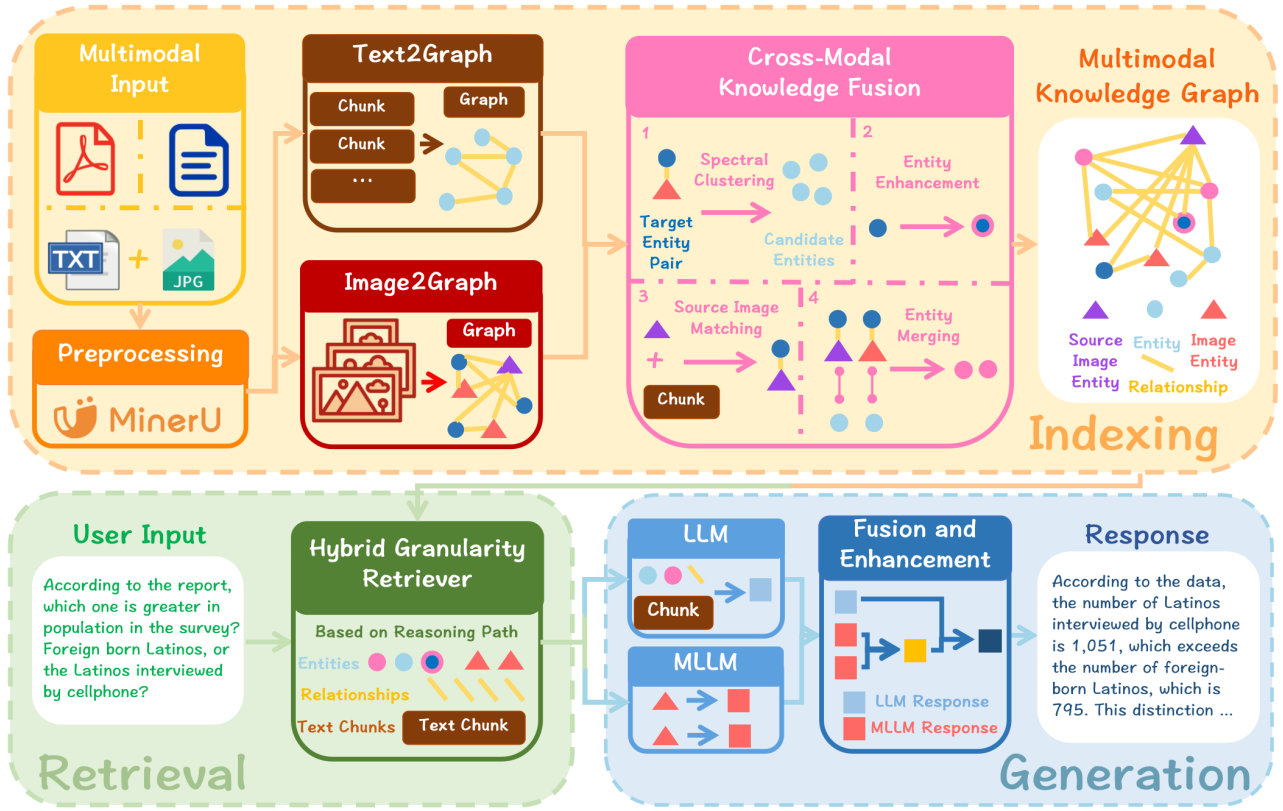


Figure 1: MMGraphRAG Framework Overview. This diagram illustrates the comprehensive workflow of the MMGraphRAG framework, starting with the input of multi-modal data. After preprocessing, Text2KG Module transform textual data into a KG, while the Image2Graph Module processes visual data into image-based KG via its scene graph. The Cross-Modal Knowledge Fusion Module integrates text-based and image-based KGs into a unified MMKG. The Hybrid Granularity Retriever enables precise retrieval based on the structure of the MMKG, further refining the model’s ability to handle complex queries. Finally, the generative process synthesizes the retrieved multi-modal clues into a coherent and enhanced response.

MMKG for cross-modal reasoning.

2. **CMEL dataset:** We build and release the CMEL dataset, specifically designed for alignment between visual and textual entities, addressing the lack of benchmarks in this area.
3. **Spectral clustering–based CMEL method:** We design a cross-modal entity alignment process, utilizing spectral clustering to efficiently generate candidate entities by integrating semantic and structural information, thereby enhancing the accuracy of the CMEL task.

Related Work

GraphRAG

To improve the performance of LLMs on complex QA tasks, researchers have integrated knowledge graphs to enhance the reasoning capabilities and interpretability of RAG systems. GraphRAG integrates local and global knowledge by constructing entity knowledge graphs and community summaries (Edge et al. 2025). Following this, various improvements have been made, such as StructRAG, ToG-2, and Sub-graphRAG, which introduce different methods for enhancing multi-hop QA and reasoning, but these methods are still

limited to textual data (Li et al. 2025; Ma et al. 2025a; Li, Miao, and Li 2025). As for multimodal data, HM-RAG proposed a hierarchical multi-agent multimodal RAG framework that coordinates decomposition agents, multi-source retrieval agents, and a decision agent to dynamically synthesize knowledge from structured, unstructured, and graph-based data (Liu et al. 2025b). However, while HM-RAG represents progress in multimodal processing, it still relies on converting multimodal content into text via MLLMs without fully capturing the cross-modal relations, leading to incomplete logical chains.

Entity Linking

Entity Linking (EL) has evolved from text-only methods to Multimodal Entity Linking (MEL), and more recently to Cross-Modal Entity Linking (CMEL), which supports cross-modal reasoning. Traditional EL methods associate textual entities with their corresponding entries in a knowledge base, but overlook non-textual information (Shen, Wang, and Han 2015; Shen et al. 2023). MEL extends EL by incorporating visual information as auxiliary attributes to enhance alignment between entities and knowledge base entries (Gan et al. 2021; Liu et al. 2024b; Song et al. 2024).

However, MEL does not establish cross-modal relations beyond these auxiliary associations, thereby limiting genuine cross-modal interaction.

CMEL goes further by treating visual content as entities—aligning visual entities with their textual counterparts—to construct MMKGs and facilitate explicit cross-modal inference (Yao et al. 2023). Research on CMEL remains in its early stages, lacking a unified theoretical framework and robust evaluation protocols. The MATE benchmark is introduced to assess CMEL performance, but its synthetic 3D scenes fall short in capturing the complexity and diversity of real-world images (Alonso et al. 2025). To bridge this gap, we construct a CMEL dataset featuring greater real-world complexity and propose a spectral clustering-based method for candidate entity generation to drive further advances in CMEL research.

Methodology

To facilitate effective cross-modal information fusion and reasoning, we propose MMGraphRAG, a modular framework, which extends the conventional GraphRAG pipeline by incorporating dedicated mechanisms for visual modality processing and integration. It enables the joint construction of MMKGs from textual and visual inputs, thereby supporting retrieval and generation based on multimodal data.

MMGraphRAG Framework

The overall architecture of the proposed MMGraphRAG framework is illustrated in Figure 1. It consists of three modular stages: **Indexing**, **Retrieval**, and **Generation**.

Indexing Stage The goal of the indexing stage is to transform raw multimodal data (text and images) into a structured MMKG. This stage comprises three sub-modules:

- **Preprocessing Module:** This module parses input documents using tools such as *MinerU* (Wang et al. 2024a), extracting and separating textual and visual content. The data are then standardized for downstream processing.
- **Single-Modal Processing Module:** For textual inputs, document chunking and entity extraction are performed to construct a text-based KG. For visual inputs, image segmentation, scene graph construction, and entity alignment are applied to generate an image-based KG.
- **Cross-Modal Fusion Module:** This module employs our spectral clustering method to identify candidate entity pairs for fusion. It then performs cross-modal entity linking to merge text-based KG and image-based KG into a unified MMKG.

Retrieval Stage Given a query, the retrieval module performs hybrid-granularity retrieval, extracting relevant entities, relations, and contextual information from the MMKG.

Generation Stage The generation module adopts a hybrid generation strategy. Initially, an LLM generates a preliminary text-based response. Subsequently, an MLLM produces several multimodal responses grounded in both visual and textual information. These multimodal responses are then consolidated by an LLM into a unified one. Finally, the LLM

integrates the textual and the multimodal response into a unified and coherent answer. This strategy effectively mitigates the reasoning limitations of current MLLMs and ensures high-quality, contextually appropriate responses.

Advantages of the Design The MMGraphRAG framework offers several notable advantages. First, by modeling images as independent nodes, the framework enhances cross-modal reasoning capabilities. This design provides strong support for sophisticated cross-modal inference tasks. Moreover, modular architecture ensures high extensibility and adaptability. Each component is independently replaceable, allowing the system to be easily adjusted or extended to accommodate additional modalities. Finally, by constructing the MMKG based on LLMs, the framework eliminates reliance on training, offering enhanced flexibility.

Image2Graph

To realize MMGraphRAG, fine-grained entity-level processing of visual data is crucial. Constructing accurate and comprehensive scene graphs plays a key role, but traditional methods often overlook fine-grained semantic details and fail to account for hidden information between objects, leading to inadequate scene graphs and reasoning bias in downstream tasks (Liu et al. 2025b,a).

In contrast, MLLM-based methods can extract entities and infer implicit relations through semantic segmentation and reasoning ability of MLLM, generating both high-precision and fine-grained scene graphs (Chen, Li, and Wang 2024; Wang et al. 2025). These methods capture both explicit spatial relations (e.g., *girl—girl holding a camera—camera*) and implicit ones (e.g., *boy—the boy and girl appear to be close, possibly friends or a couple—girl*). Furthermore, they provide richer semantic descriptions for visual entities, refining basic labels such as *boy* into more detailed expressions like *a college student with tired eyes*. Unlike methods relying on large-scale annotated datasets (Lin et al. 2020; Suhail et al. 2021; Tang et al. 2020; Xu et al. 2017; Zellers et al. 2018; Zheng et al. 2023), MLLM-based methods improve generalization capabilities by minimizing human supervision.

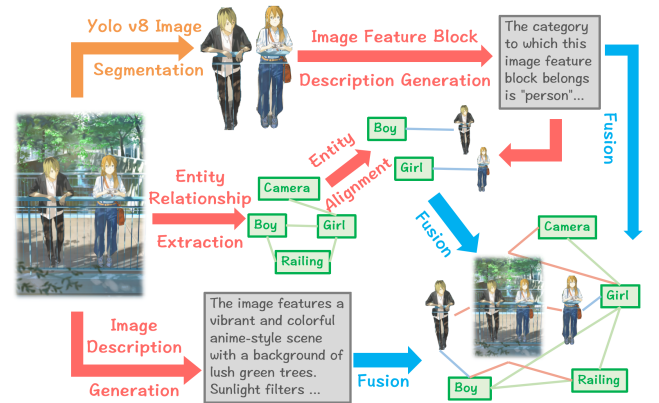


Figure 2: An Example of the Img2Graph Module in Action

The Img2Graph module maps images into knowledge

graphs through a five-step pipeline. First, semantic segmentation is performed using YOLO (Ultralytics 2023) to divide the image into semantically independent regions, referred to as image feature blocks. Second, MLLMs generate textual descriptions for each feature block. Third, entities and their relations are extracted from the image. Fourth, MLLMs align the segmented feature blocks with the extracted entities. Fifth, a global entity is constructed to describe the entire image and to establish connections with local entities. A detailed example is illustrated in Figure 2.

Through this pipeline, Img2Graph produces scene graphs with explicit structure and enriched semantics, transforming raw visual inputs into KGs. Each feature block possesses clear physical boundaries and, with the aid of MLLMs, is assigned accurate and informative descriptions. The resulting multimodal scene graphs not only enhance the structural representation of visual content but also strengthen contextual relevance in multimodal semantic retrieval. This extends GraphRAG with improved retrievability and clearer reasoning paths, substantially enhancing the robustness and accuracy of the RAG process in complex query scenarios.

Cross-modal Fusion Method

The core objective of cross-modal fusion is to construct a unified MMKG by aligning and fusing entities from both image-based and text-based KGs. This process not only ensures deep semantic alignment between modalities, but also significantly enhances the logical coherence and informational completeness of the resulting MMKG. Specifically, our cross-modal fusion framework consists of the following key components:

1. Fine-Grained Entity Alignment Between KGs The first and most crucial step of cross-modal fusion is aligning entities extracted from image and text modalities, which is essentially Cross-Modal Entity Linking (CMEL).

CMEL Task Definition Let the image set be defined as $I = \{I_1, I_2, \dots, I_N\}$, where each image I_i contains a set of extracted entities denoted by: $E(I_i) = \{e_1^{(I_i)}, e_2^{(I_i)}, \dots, e_K^{(I_i)}\}$.

Similarly, define the text set as $T = \{T_1, T_2, \dots, T_M\}$, where each text chunk T_j contains a set of extracted entities: $E(T_j) = \{e_1^{(T_j)}, e_2^{(T_j)}, \dots, e_L^{(T_j)}\}$.

The goal of CMEL is to identify pairs of entities from images and text that refer to the same real-world concept. In other words, for each visual entity $e_k^{(I_i)}$, we aim to align it with the most semantically relevant textual entity $e_l^{(T_j)}$. Since the number of textual entities is generally larger than that of visual entities, the task is decomposed into two stages: (1) generating a set of candidate textual entities for each visual entity, and (2) selecting the best-aligned textual entity from this set.

Formally, for each visual entity $e_k^{(I_i)}$, we define the candidate set as:

$$C(e_k^{(I_i)}) \subseteq \bigcup_{j=1}^{j+1} E(T_j),$$

where $C(e_k^{(I_i)})$ contains the most relevant textual entities selected from the textual entity pool of the context.

The final alignment is determined by maximizing a similarity function f , such that:

$$\mathcal{A}(e_k^{(I_i)}) = \arg \max_{e \in C(e_k^{(I_i)})} f(e_k^{(I_i)}, e).$$

Once aligned, the linked entity pairs are passed to an LLM-based fusion module to ensure they share a unified representation in the KG.

Spectral Clustering-Based Candidate Generation To improve the efficiency and robustness of candidate entities generation, we propose a spectral clustering-based optimization strategy. Existing methods fall into two categories: (1) *distance-based clustering*, such as KMeans (Kodinariya, Makwana et al. 2013; Likas, Vlassis, and J. Verbeek 2003; Sinaga and Yang 2020) and DBSCAN (Deng 2020; Khan et al. 2014), which depends semantic similarity but ignores graph structure, and (2) *graph-based clustering*, such as PageRank (A. Tabrizi et al. 2013) and Leiden (Traag, Waltman, and Van Eck 2019), which captures structural relations but suffers in sparse graphs. To address both aspects, we design a spectral clustering algorithm tailored for CMEL.

Specifically, we redesign the weighted adjacency matrix \mathbf{A} and the degree matrix \mathbf{D} to capture both semantic and structural information between entities.

The adjacency matrix \mathbf{A} is constructed to reflect the similarity between nodes as well as the importance of their relations. It is defined as:

$$A_{pq} = \text{sim}(\mathbf{v}_p, \mathbf{v}_q) \cdot \text{weight}(r_{pq}) \quad (1)$$

where \mathbf{v}_p is the embedding vector of entity e_p , and $\text{sim}(\cdot)$ denotes cosine similarity. The term r_{pq} represents the relation between e_p and e_q in the KG, and $\text{weight}(r_{pq})$ is a scalar reflecting the importance of the relation assessed by LLMs. If no relation exists between two entities, we set $\text{weight}(r_{pq}) = 1$.

The degree matrix \mathbf{D} is a diagonal matrix, where each diagonal entry D_{pp} indicates the connectivity strength of node p , computed as:

$$D_{pp} = \sum_q A_{pq}.$$

Intuitively, each diagonal value in D represents the total weighted similarity between node p and all other nodes.

Following the standard spectral clustering procedure, we construct the Laplacian matrix and perform eigen-decomposition. We then form the matrix $\mathbf{Q} = [\mathbf{u}_1, \dots, \mathbf{u}_m]$ using the smallest m eigenvectors, where m depends on the number of textual entities in context (Jia et al. 2014).

Clustering is performed on the row space of \mathbf{Q} using DBSCAN to obtain cluster partitions:

$$\text{Cluster}(\mathbf{Q}) = \{C_1, C_2, \dots, C_n\}.$$

For each image entity $e_k^{(I_i)}$, we select the most relevant cluster based on the cosine similarity between its embedding

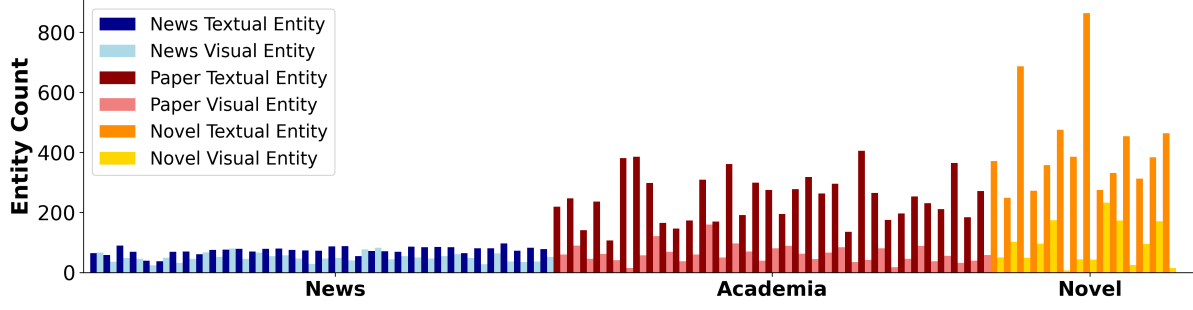


Figure 3: Entity Distribution Across Document Domains.

$\mathbf{v}_k^{(I_i)}$ and the cluster members. The candidate entity set is then defined as:

$$C(e_k^{(I_i)}) = \bigcup_{C_n} \{e_p \mid e_p \in C_n\}.$$

Finally, we perform entity alignment using LLM-based inference, which has demonstrated high accuracy and adaptability in complex alignment scenarios (Liu et al. 2024b,c). The prompt includes:

- the name and description of the visual entity,
- descriptions of candidate entities from the selected cluster, and
- a fixed set of alignment examples.

The output is adopted as the final alignment result.

2. Enhancing Remaining Image Entities Then, we enhance the descriptions of remaining visual entities not aligned during CMEL. This step incorporates semantically relevant information from the original text to improve the completeness of the image-based KG.

For instance, visual entities such as *a flooded neighborhood* can be enriched using textual entities like *Hurricane Ian* and *Florida*, yielding a more informative description: *A neighborhood in Florida severely flooded by Hurricane Ian, causing significant damage and displacement*. This step ensures high-fidelity entity descriptions and improves the internal completeness of the MMKG.

3. Aligning Global Image Entity with Relevant Textual Entity In addition to fine-grained alignment, we align the global entity of each image with relevant textual entity. If no direct match is found, we create a new entity in the text-based KG to represent the overall semantic content of the image. This step ensures that holistic semantic information is preserved and cross-modal coherence is maintained.

4. Entity Fusion for Unified Representation For all aligned entities, we perform semantic fusion to ensure unified representations in the MMKG. This step guarantees consistent alignment and representation of entities across modalities, facilitating downstream reasoning and retrieval.

5. Iterative Graph Construction The above steps are repeated for each image to construct a thorough MMKG.

Experiments

In this section, we conduct experiments on CMEL task to validate the effectiveness of the spectral clustering-based candidate entities generation method in complex multimodal scenarios. We then evaluate the overall performance of the MMGraphRAG framework on two multimodal Document Question Answering (DocQA) benchmarks, DocBench and MMLongBench, followed by an analysis of advantages.

CMEL Experiments

To evaluate the effectiveness of our proposed spectral clustering-based method for the CMEL task, we construct a novel CMEL benchmark. Designed for fine-grained multi-entity alignment in complex multimodal scenarios, the CMEL dataset exhibits significantly greater entity diversity and relation complexity compared to existing benchmarks such as MATE. Moreover, CMEL dataset is released as an open-source benchmark and supports extensibility through a semi-automated construction pipeline. This pipeline enables users to incorporate minimal human supervision to generate new samples, providing a sustainable experimental foundation for future research in the CMEL task.

CMEL Dataset The CMEL dataset comprises documents from three distinct domains—news, academia, and novels—ensuring broad domain diversity and practical applicability, illustrated in Figure 3. Each sample includes (i) a text-based KG built from text chunks, (ii) an image-based KG derived from per-image scene graphs, and (iii) the original PDF-format document. In total, CMEL provides 1,114 alignment instances—87 drawn from news articles, 475 from academic papers, and 552 from novels.

To comprehensively assess performance, we adopt both micro-accuracy and macro-accuracy as evaluation metrics. Micro-accuracy is computed on a per-entity basis, reflecting the overall prediction correctness and serving as an indicator of global performance. Meanwhile, macro-accuracy calculates the average accuracy per document, mitigating evaluation bias caused by imbalanced entity distributions across documents and better highlighting performance of different methods across diverse domains.

Experimental Setup and Results We conduct a series of comparative experiments based on the CMEL dataset. The experiments cover three categories of approaches:

embedding-based methods (Emb), LLM-based methods (LLM), and our proposed spectral clustering-based candidate entities generation method (Spec). We further compare against multiple mainstream clustering algorithms to provide comprehensive baselines.

The embedding-based method encodes visual and textual entities into vector representations using a pretrained embedding model, and then computes their semantic proximity via cosine similarity. A textual entity is considered a candidate if the similarity score exceeds a predefined threshold.

The LLM-based method leverages the reasoning capabilities of LLMs to directly generate candidate sets. Specifically, the model is prompted with an visual entity, its surrounding context, and a pool of textual entities. It then outputs the most plausible candidate entities based on contextual understanding.

For clustering-based baselines, we include DBSCAN (DB), KMeans (KM), PageRank (PR), and Leiden (Lei).

Entity alignment within candidate sets is uniformly conducted via LLM-based reasoning.

To ensure the robustness and generalizability of our results, evaluations are conducted on different models. Due to space limitations, we report only the best-performing configurations for each method category in the results table. Specifically, the experiments utilize the embedding model stella-en-1.5B-v5 (AI 2024c), the LLM Qwen2.5-72B-Instruct (Team 2024), and the MLLM InternVL2.5-38B-MPO (Wang et al. 2024c).

Meth.	micro/macro Acc.			Overall.
	News	Aca.	Nov.	
Emb	10.8/8.4	33.1/34.5	9.0/7.5	20.0/16.8
LLM	33.3/24.1	36.8/36.1	17.4/20.8	27.1/27.0
DB	53.8/45.9	60.8/58.3	29.9/34.2	45.2/46.1
KM	50.5/40.6	60.7/57.7	29.6/30.5	45.2/43.0
PR	51.6/44.4	59.7/56.8	29.1/35.2	44.1/45.5
Lei	54.8/44.7	60.5/55.5	29.4/30.6	44.8/43.6
Spec	65.5/56.9	73.3/69.9	31.2/39.4	51.8/59.2

Table 1: Micro/macro accuracy on CMEL dataset

The results are shown in Table 1. Overall, clustering-based methods significantly outperform the embedding-based and LLM-based methods in the CMEL task. Compared to other clustering-based methods, our spectral clustering-based method performs best and improves micro-accuracy by about 15% and macro-accuracy by around 30%, clearly demonstrating the effectiveness of spectral clustering-based candidate entities generation method.

Multimodal Document QA Experiments

We choose DocQA as the primary evaluation task for MM-GraphRAG because it comprehensively assesses the capabilities of method in multimodal information integration, complex reasoning, and domain adaptability. DocQA task involves deep understanding of long documents and the integration of diverse formats. Furthermore, documents from

different domains exhibit unique terminologies and structural patterns, requiring methods to adapt flexibly. These characteristics make multimodal DocQA a challenging and comprehensive benchmark for evaluating the performance of MMGraphRAG.

Benchmarks DocBench contains 229 PDF documents from publicly available online resources, covering five domains: academia (Aca.), finance (Fin.), government (Gov.), laws (Law.), and news (News). It includes four types of questions: pure text questions (Txt.), multimodal questions (Mm.), metadata questions, and unanswerable questions (Una.). For our experiments, since the information is converted into KGs, we do not focus on metadata. Therefore, this category of questions is excluded from statistics. For evaluation, DocBench determines the correctness of answers using LLM (Llama3.1-70B-Instruct in the experiments).

MMLongBench consists of 135 long PDF documents from seven different domains. MMLongBench includes annotations spanning multiple sources of evidence, such as text (Txt.), charts, tables (C.T.), layout (Lay.), and figures (Fig.). MMLongBench follows the three-step evaluation protocol of MATHVISTA (Lu et al. 2023): response generation, answer extraction using LLM, and score calculation. Accuracy and F1 scores are reported to balance the evaluation of answerable and unanswerable questions, using Llama3.1-70B-Instruct throughout.

Comparison with Basic Methods To ensure fair comparisons and eliminate potential biases arising from using the same model for both generation and evaluation, this experiment selects various LLMs and MLLMs as comparison benchmarks. Although our experiments initially evaluated three LLMs, such as Llama3.1-70B-Instruct (AI 2024a), Qwen2.5-72B-Instruct (Yang et al. 2024), and Mistral-Large-Instruct-2411 (AI 2024b) and three MLLMs, such as Ovis1.6-Gemma2-27B (Lu et al. 2024), Qwen2-VL-72B (Wang et al. 2024b), and InternVL2.5-38B-MPO (Chen et al. 2024b), we choose the two most representative models to make it more clear: Qwen2.5-72B-Instruct for the LLM comparison, and InternVL2.5-38B-MPO for the MLLM comparison.

We consider the following baselines:

LLM: We replace images with MLLM generated corresponding text after reprocessing. All text and questions are input into the LLM. If the content exceeds the model’s context length, it will be divided into parts, and partial answers are concatenated as final results.

MLLM: All images are concatenated and resized based on model constraints. These image blocks, along with the question, are input into the MLLM to test its ability to reason over multimodal data.

NaiveRAG (Lewis et al. 2020b) (NRAG): The text is chunked into segments of 500 tokens. Each chunk and question is embedded. Then, the top-k relevant chunks (10 selected) are retrieved by calculating cosine similarity with the question, and these relevant chunks are provided along with the question to LLMs.

GraphRAG (GRAG): The GraphRAG method was modified by removing the community detection part and using

local mode querying to ensure fair comparison with other methods (Guo et al. 2024; Edge et al. 2024). The top-k entities (10 selected) are used for retrieval, with the length of the text of entities and relations limited to a maximum of 4000 tokens, and the maximum number of chunks is 10.

We compared multiple methods and highlighted the advantages of MMGraphRAG (MMGR) over other methods.

Meth.	Types					Domains			Acc.
	Aca.	Fin.	Gov.	Law.	News	Txt.	Mm.	Una.	
LLM	41.3	16.3	50.7	49.7	77.3	53.9	20.1	75.8	44.8
MLLM	19.8	16.3	28.4	31.4	46.5	35.7	15.9	39.5	27.7
NRAG	43.6	34.4	62.8	65.4	75.0	81.6	30.5	67.7	59.5
GRAG	39.6	25.7	52.5	49.7	74.5	71.7	26.0	67.5	52.3
MMGR	60.5	65.8	66.5	70.4	<u>77.1</u>	<u>81.2</u>	88.7	<u>71.9</u>	76.8

Table 2: Accuracy on DocBench

Meth.	Locations			Formats				Overall	
	Sin.	Mul.	Una.	C.T.	Txt.	Lay.	Fig.	Acc.	F1
LLM	22.5	20.0	53.2	14.7	33.3	23.5	16.1	27.8	22.1
MLLM	13.3	7.9	13.9	8.4	10.7	11.8	11.4	11.6	10.4
NRAG	22.3	16.4	52.5	13.0	30.3	20.3	14.9	26.2	20.9
GRAG	18.2	13.2	77.1	12.5	26.1	12.2	8.5	<u>28.1</u>	19.3
MMGR	39.6	26.7	<u>55.8</u>	35.6	33.8	28.7	34.6	38.8	34.1

Table 3: Accuracy and F1 score on MMLongBench

The results of the DocBench dataset are shown in Table 2, and the results of the MMLongBench dataset are shown in Table 3. The following analysis from several key perspectives explains the superiority of MMGraphRAG.

Multimodal Information Processing Capability Counterintuitively, results show that MLLMs alone do not outperform other methods on vision-centric queries, even worse than NaiveRAG. MMGraphRAG improves this by constructing fine-grained image-based KGs, enabling more precise retrieval of visual content relevant to the question. This allows MMGraphRAG to better exploit the visual understanding capabilities of MLLMs, thereby improving performance on vision-centric queries.

Multimodal Fusion and Reasoning MMGraphRAG achieves superior results on both textual and multimodal questions compared to GraphRAG (MMGR 81.2 vs. GRAG 71.7; MMGR 88.7 vs. GRAG 26.0 on DocBench), validating the importance of cross-modal fusion for complex reasoning. For example, when answering questions spanning multiple pages, MMGraphRAG can follow reasoning paths across pages and modalities via the MMKG, synthesizing distributed evidence into coherent answers.

Cross-Domain Adaptability Compared to all text-only RAG methods, MMGraphRAG shows substantial gains in domains with high visual-structural complexity, such as academia and finance. This indicates that MMGraphRAG

Meth.	Locations			Formats				Overall	
	Sin.	Mul.	Una.	C.T.	Txt.	Lay.	Fig.	Acc.	F1
M3DR	32.4	14.8	5.8	39.0	30.0	23.5	20.8	21.0	22.6
MMGR	34.3	12.5	35.1	48.2	24.6	18.2	22.2	26.5	23.8

Table 4: M3DOCRAG vs MMGraphRAG performance

performs well in specialized fields while retaining the ability to generalize across diverse domains. Its flexibility and adaptability make it well-suited for real-world applications involving heterogeneous multimodal documents.

Comparison with MRAG Methods To further validate the effectiveness and advantages of the MMGraphRAG framework, we conduct a direct comparison with M3DOCRAG (M3DR), one of the state-of-the-art methods for multimodal DocQA. All experiments were conducted on the MMLongBench dataset, using Qwen2.5-7B-Instruct and Qwen2-VL-7B. Across multiple evaluation dimensions, MMGraphRAG demonstrates significantly superior performance in multimodal information understanding, primarily attributed to its novel mechanisms for MMKG construction. Detailed results are shown in Table 4.

By explicitly linking visual and textual entities within the MMKG, MMGraphRAG enables more efficient handling of complex cross-modal reasoning tasks. It consistently outperforms M3DOCRAG in scenarios that involve visual content such as charts, tables, and figures. Unlike DocBench, MMLongBench does not isolate metadata questions; hence M3DOCRAG performs well when the answer hinges on text or layout cues, yet it struggles in cases requiring deep semantic alignment between images and text—particularly in table-related queries (MMGR 48.2 vs. M3DR 39.0).

Moreover, MMGraphRAG exhibits a clear advantage in handling unanswerable questions. Existing MRAG methods typically rely on large-scale training data and often lack dedicated mechanisms or negative samples tailored for unanswerable cases. Additionally, due to projecting textual, visual, and layout features into a shared embedding space, even slightly related but noisy evidence can mislead the model into confidently generating incorrect answers.

In contrast, MMGraphRAG achieves more complete and fine-grained cross-modal information interactions through CMEL. Structured reasoning is then conducted over the resulting MMKG, enabling the model to more reliably assess whether a question is answerable. This reduces the generation of misleading answers and enhances robustness in real-world multimodal QA scenarios.

Conclusion

We propose MMGraphRAG, a framework that integrates textual and image data into a multimodal knowledge graph to facilitate deep cross-modal fusion and reasoning. Experimental results demonstrate that MMGraphRAG outperforms all kinds of existing RAG methods on multimodal DocQA tasks. The framework exhibits strong domain adaptability and produces traceable reasoning paths. We hope this work

will inspire further research on cross-modal entity linking and the development of graph-based frameworks for deeper and more comprehensive multimodal reasoning.

References

- A. Tabrizi, S.; Shakery, A.; Asadpour, M.; Abbasi, M.; and Tavallaie, M. A. 2013. Personalized PageRank Clustering: A graph clustering algorithm based on random walks. *Physica A: Statistical Mechanics and its Applications*, 392(22): 5772–5785.
- AI, M. 2024a. LLaMA-3.1-70B-Instruct. Available at <https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>.
- AI, M. 2024b. Mistral-Large-Instruct-2411. Available at <https://huggingface.co/mistralai/Mistral-Large-Instruct-2411>.
- AI, S. 2024c. stella-en-1.5B-v5. <https://huggingface.co/stabilityai/stella-en-1.5B-v5>. Open-weight English language model.
- Alonso, I.; Azkune, G.; Salaberria, A.; Barnes, J.; and de Lacalle, O. L. 2025. Vision-Language Models Struggle to Align Entities across Modalities. *arXiv:2503.03854*.
- Chen, G.; Li, J.; and Wang, W. 2024. Scene Graph Generation with Role-Playing Large Language Models. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 132238–132266. Curran Associates, Inc.
- Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024a. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Chen, Y.; Sun, P.; Li, X.; and Chu, X. 2025. MRD-RAG: Enhancing Medical Diagnosis with Multi-Round Retrieval-Augmented Generation. *arXiv:2504.07724*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. 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*, 24185–24198.
- Cho, J.; Mahata, D.; Irsoy, O.; He, Y.; and Bansal, M. 2024. M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. *CoRR*, abs/2411.04952.
- Chu, Y.-W.; Zhang, K.; Malon, C.; and Min, M. R. 2025. Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.
- Deng, D. 2020. DBSCAN Clustering Algorithm Based on Density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 949–953.
- Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv:2404.16130*.
- Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; HUDELLOT, C.; and Colombo, P. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations*.
- Gan, J.; Luo, J.; Wang, H.; Wang, S.; He, W.; and Huang, Q. 2021. Multimodal Entity Linking: A New Dataset and A Baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM ’21, 993–1001. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.
- Guo, G.; Wang, H.; Bell, D.; Bi, Y.; and Greer, K. 2003. KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, 986–996. Springer.
- Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; and Huang, C. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.
- Hindi, M.; Mohammed, L.; Maaz, O.; and Alwarafy, A. 2025. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access*, 13: 46171–46189.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2).
- Jia, H.; Ding, S.; Xu, X.; and Nie, R. 2014. The latest research progress on spectral clustering. *Neural Computing and Applications*, 24: 1477–1486.
- Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. Ultralytics YOLO v8. Available at <https://github.com/ultralytics/ultralytics>.
- Khan, K.; Rehman, S. U.; Aziz, K.; Fong, S.; and Sarasvady, S. 2014. DBSCAN: Past, present and future. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 232–238.
- Kodinariya, T. M.; Makwana, P. R.; et al. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal*, 1(6): 90–95.
- Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020a. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and

- Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.
- Li, M.; Miao, S.; and Li, P. 2025. Simple is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation. In *The Thirteenth International Conference on Learning Representations*.
- Li, Z.; Chen, X.; Yu, H.; Lin, H.; Lu, Y.; Tang, Q.; Huang, F.; Han, X.; Sun, L.; and Li, Y. 2025. StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization. In *The Thirteenth International Conference on Learning Representations*.
- Liang, W.; Meo, P. D.; Tang, Y.; and Zhu, J. 2024. A Survey of Multi-modal Knowledge Graphs: Technologies and Trends. *ACM Comput. Surv.*, 56(11).
- Likas, A.; Vlassis, N.; and J. Verbeek, J. 2003. The global k-means clustering algorithm. *Pattern Recognition*, 36(2): 451–461. Biometrics.
- Lin, W.; Chen, J.; Mei, J.; Coca, A.; and Byrne, B. 2023. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 22820–22840. Curran Associates, Inc.
- Lin, X.; Ding, C.; Zeng, J.; and Tao, D. 2020. GPS-Net: Graph Property Sensing Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ling, Z.; Guo, Z.; Huang, Y.; An, Y.; Xiao, S.; Lan, J.; Zhu, X.; and Zheng, B. 2025. MMKB-RAG: A Multi-Modal Knowledge-Based Retrieval-Augmented Generation Framework. *arXiv:2504.10074*.
- Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024a. A Survey on Hallucination in Large Vision-Language Models. *CoRR*, abs/2402.00253.
- Liu, J.; Meng, S.; Gao, Y.; Mao, S.; Cai, P.; Yan, G.; Chen, Y.; Bian, Z.; Shi, B.; and Wang, D. 2025a. Aligning Vision to Language: Text-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning. *arXiv:2503.12972*.
- Liu, P.; Liu, X.; Yao, R.; Liu, J.; Meng, S.; Wang, D.; and Ma, J. 2025b. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. *arXiv:2504.12330*.
- Liu, Q.; He, Y.; Xu, T.; Lian, D.; Liu, C.; Zheng, Z.; and Chen, E. 2024b. UniMEL: A Unified Framework for Multimodal Entity Linking with Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, 1909–1919. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704369.
- Liu, X.; Liu, Y.; Zhang, K.; Wang, K.; Liu, Q.; and Chen, E. 2024c. OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting. *arXiv preprint arXiv:2410.07549*.
- Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.
- Ma, S.; Xu, C.; Jiang, X.; Li, M.; Qu, H.; Yang, C.; Mao, J.; and Guo, J. 2025a. Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation. In *The Thirteenth International Conference on Learning Representations*.
- Ma, Y.; Zang, Y.; Chen, L.; Chen, M.; Jiao, Y.; Li, X.; Lu, X.; Liu, Z.; Ma, Y.; Dong, X.; et al. 2025b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37: 95963–96010.
- Mei, L.; Mo, S.; Yang, Z.; and Chen, C. 2025. A Survey of Multimodal Retrieval-Augmented Generation. *arXiv:2504.08748*.
- Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2020. Recipes for building an open-domain chatbot. *CoRR*, abs/2004.13637.
- Shen, W.; Li, Y.; Liu, Y.; Han, J.; Wang, J.; and Yuan, X. 2023. Entity Linking Meets Deep Learning: Techniques and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 35(3): 2556–2578.
- Shen, W.; Wang, J.; and Han, J. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2): 443–460.
- Sinaga, K. P.; and Yang, M.-S. 2020. Unsupervised K-Means Clustering Algorithm. *IEEE Access*, 8: 80716–80727.
- Song, S.; Zhao, S.; Wang, C.; Yan, T.; Li, S.; Mao, X.; and Wang, M. 2024. A Dual-Way Enhanced Framework from Text Matching Point of View for Multimodal Entity Linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 19008–19016.
- Suhail, M.; Mittal, A.; Siddiquie, B.; Broaddus, C.; Eleath, J.; Medioni, G.; and Sigal, L. 2021. Energy-Based Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13936–13945.
- Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation From Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Team, Q. 2024. Qwen2.5: A Party of Foundation Models. <https://qwenlm.github.io/blog/qwen2.5/>. Includes instruction-tuned Qwen2.5-72B-Instruct via Hugging Face.
- Traag, V. A.; Waltman, L.; and Van Eck, N. J. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1): 1–12.
- Ultralytics. 2023. Ultralytics YOLOv8: Cutting-Edge Object Detection Models. <https://github.com/ultralytics/ultralytics/>. Accessed: 2025-07-14.
- Wang, B.; Xu, C.; Zhao, X.; Ouyang, L.; Wu, F.; Zhao, Z.; Xu, R.; Liu, K.; Qu, Y.; Shang, F.; et al. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.
- Wang, J.; Ju, J.; Luan, J.; and Deng, Z. 2025. LLaVA-SG: Leveraging Scene Graphs as Visual Semantic Expression in Vision-Language Models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; Dai, J.; and et al. 2024c. InternVL2.5-MPO: Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv preprint arXiv:2411.10442*.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.
- Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2025. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *The Thirteenth International Conference on Learning Representations*.
- Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yao, B. M.; Chen, Y.; Wang, Q.; Wang, S.; Liu, M.; Xu, Z.; Yu, L.; and Huang, L. 2023. AMELI: Enhancing Multimodal Entity Linking with Fine-Grained Attributes. *CoRR*, abs/2305.14725.
- Yin, C.; and Zhang, Z. 2024. A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With “Same Semantics, Different Structure” After Fine Tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, 677–684. Atlantis Press.
- Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; and Sun, M. 2024. VisRAG: Vision-based Retrieval-augmented Generation on Multimodality Documents. *CoRR*, abs/2410.10594.
- Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural Motifs: Scene Graph Parsing With Global Context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhang, D.; Li, J.; Zeng, Z.; and Wang, F. 2024. Jasper and Stella: distillation of SOTA embedding models. *arXiv preprint arXiv:2412.19048*.
- Zheng, C.; Lyu, X.; Gao, L.; Dai, B.; and Song, J. 2023. Prototype-Based Embedding Network for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22783–22792.
- Zou, A.; Yu, W.; Zhang, H.; Ma, K.; Cai, D.; Zhang, Z.; Zhao, H.; and Yu, D. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *arXiv preprint arXiv:2407.10701*.

MMGraphRAG: Bridging Vision and Language with Interpretable Multimodal Knowledge Graphs

Supplementary Material

CMEL dataset

The CMEL dataset is a novel benchmark designed to facilitate the evaluation of Cross-Modal Entity Linking (CMEL) tasks, focusing on fine-grained cross-entity alignment in complex multimodal scenarios. It features greater entity diversity and relational complexity compared to existing datasets like MATE. The dataset comprises documents from three distinct domains—news, academia, and novels—ensuring broad domain diversity. Each sample includes a text-based knowledge graph (KG) derived from text chunks, an image-based KG created from per-image scene graphs, and the original PDF-format document. In total, the CMEL dataset includes 1,114 alignment instances, divided into 87 from news articles, 475 from academic papers, and 552 from novels. A more detailed introduction is conducted in Subsection A. Additionally, the dataset supports extensibility via a semi-automated construction pipeline, allowing for minimal human supervision to generate new samples. The construction process is introduced in Subsection B. And in Subsection C, the full experiment results on the CMEL dataset are presented.

A Detailed Introduction to the CMEL Dataset

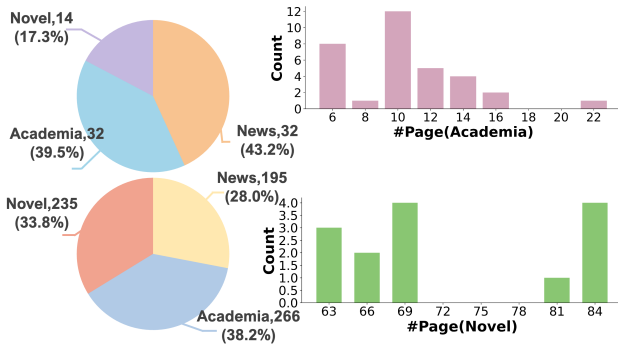


Figure 4: The Distribution of CMEL dataset. In the top-left, the number and proportion of documents in each domain are shown; in the bottom-left, the number and proportion of images in each domain are displayed; in the top-right, the page distribution of academia domain documents is provided; and in the bottom-right, the page distribution of novel documents is shown. All news domain documents are one page.

The distribution of dataset documents and images is shown in Figure 4. The entire dataset is constructed based on the number of images, so when divided by images, the number of documents in the three domains is approximately equal.

In addition to the text and image knowledge graphs, the CMEL dataset also contains a wealth of supplementary information to assist with the CMEL task. The text in the original documents is extracted into Markdown format using MinerU(Wang et al. 2024a) and stored in the form of text chunks. The image information includes various details such as the corresponding text chunks and image descriptions, all summarized in a JSON file named `kv_store_image_data`. Take the entity corresponding to Figure 5 as a specific example.

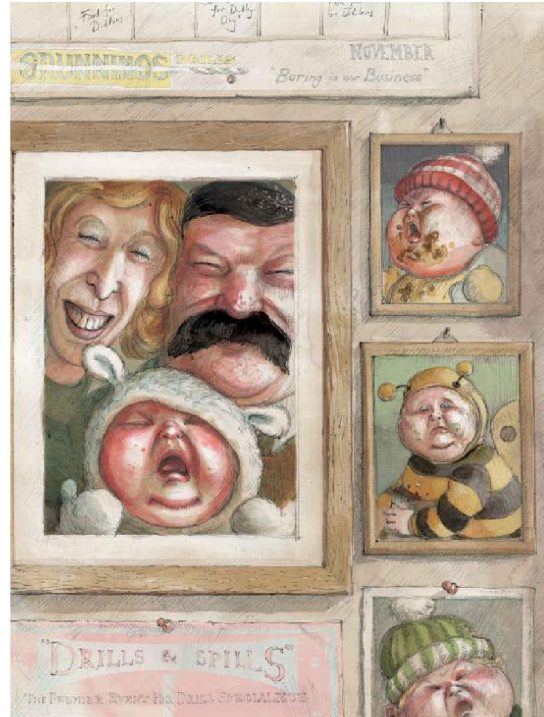


Figure 5: The Dursleys' photo wall. From Chapter 1 of "Harry Potter and the Sorcerer's Stone".

The original images are numbered starting from 1 in the order of their appearance and stored separately. The storage location can also be found in `kv_store_image_data`. Example data is as follows:

```

"image_2": {
  "image_id": 2,
  "image_path": "./images/image_2.jpg"
  ,
  "caption": [],
  "footnote": [],
  "context": "Mr. Dursley was the
    director of a firm called
    Grunnings...",
  "chunk_order_index": 0,
  "chunk_id": "chunk-
    fb8e5b95ca964e204d9e59caeaf25f09
    ",
  "description": "The image depicts a
    wall adorned with framed
    pictures and posters. The
    central frame contains a family
    portrait featuring two adults
    and a baby...",
  "segmentation": true
}

```

The ground truth for each instance is stored in a JSON file. Example data is as follows:

```

"image_2": [
  {
    "merged_entity_name": "BABY IN
      RED HAT",
    "entity_type": "PERSON",
    "description": "A small framed
      picture of a baby wearing a
      red hat with a sad
      expression...",
    "source_image_entities": [
      "BABY IN RED HAT"
    ],
    "source_text_entities": [
      "DUDLEY"
    ]
  },
  ...
]

```

The calculation formulas for the micro and macro accuracy in the CMEL dataset are as follows:

$$\text{Micro - Accuracy} = \frac{\sum_{i=1}^N \text{Correct}_i}{\sum_{i=1}^N \text{Total}_i} \quad (2)$$

Where N is the total number of entities. Correct_i is the number of correct predictions for the i^{th} image or entity, and Total_i is the total number of predictions for the i^{th} image or entity.

$$\text{Macro - Accuracy} = \frac{1}{M} \sum_{j=1}^M \frac{\text{Correct}_j}{\text{Total}_j} \quad (3)$$

Where M is the total number of documents. Correct_j is the number of correct predictions for the j^{th} document, and Total_j is the total number of predictions for the j^{th} document.

Construction of the CMEL Dataset

Step 0: Document Collection. The documents for the news and academia domains in the CMEL dataset are sourced from the DocBench dataset(Zou et al. 2024). As for the novel domain, we choose several works especially suitable for the CMEL task. Specifically:

- In the academia domain, the papers come from arXiv, focusing on the top-k most cited papers in the natural language processing field on Google Scholar.
- In the news domain, the documents are collected from the front page scans of The New York Times, covering dates from February 22, 2022, to February 22, 2024.
- For the novel domain, four novels with a large number of images were downloaded from Zlibrary. To facilitate knowledge graph construction and manual inspection, these novels were split into 14 documents with approximately the same number of pages.

Step 1: Indexing. In this step, we follow the process introduced in Methodology to construct the initial knowledge graphs for the raw documents, including both text-based and image-based knowledge graphs. The specific operations are as follows:

- **Text-Based Knowledge Graph Construction:** First, text information is extracted from the raw PDF documents and chunked (fixed token sizes). Each text chunk is converted into a knowledge graph using LLMs, and stored in the file `kv_store_chunk_knowledge_graph.json`.
- **Image-Based Knowledge Graph Construction:** An independent image knowledge graph is constructed and stored in the file `kv_store_image_knowledge_graph.json`. Each image is linked to a scene graph, and its associated entity information, such as entity name, type, and description, is extracted and stored in the file `kv_store_image_data.json`.
- **Data Cleaning and Preparation:** Before storing the data, the working directory is cleaned, retaining only necessary files to ensure the cleanliness of the data storage.

Step 2: Check 1. In this step, LLM is used to determine whether there are any duplicate entities between different text chunks, assisting with manual inspection and corrections. The specific operations are as follows:

- **Adjacency Entities Extraction:** The `get_all_neighbors` function is used to extract adjacent entities associated with each text chunk to identify potential duplicate entities.
- **Entity Merge Prompt Generation:** Based on the content and entities of each text chunk, generate specific prompt. Then utilize LLM to determine whether these entities might be duplicates and provide suggestions for merging.
- **Manual Inspection:** The results from the LLM are manually reviewed to identify any duplicate entities and to edit the `merged_entities.json`, which serves as guides for the next step.

The prompt for entity merging is as follows:

Prompt for Finding Duplicate Entities

You are an information processing expert tasked with determining whether multiple entities represent the same object and merging the results. Below are the steps for your task:

1. You will receive a passage of text, a list of entities extracted from the text, and each entity's corresponding type and description.

2. Your task is to determine, based on the entity names, types, descriptions, and their contextual relationships in the text, which entities actually refer to the same object.

3. If you identify two or more entities as referring to the same object, merge them into a unified entity record:

entity_name: Use the most common or universal name (if there are aliases, include them in parentheses). entity_type: Ensure category consistency. description: Combine the descriptions of all entities into a concise and accurate summary. source_entities: Include all entities that were merged into this entity record.

4. The output should contain only merged entities—entities that represent the same object and have been merged. Do not include any entity records for entities that were not merged.

-Input-

Passage:

A passage providing contextual information will be given here.

Entity List:

```
[{"entity_name": "Entity1", "entity_type": "Category1", "description": "Description1", "entity_name": "Entity2", "entity_type": "Category2", "description": "Description2", ...}]
```

-Output-

```
[{"entity_name": "Unified Entity Name", "entity_type": "Unified Category", "description": "Combined Description", "source_entities": ["Entity1", "Entity2"], ...}]
```

-Considerations for Judgment-

1. Name Similarity: Whether the entity names are identical, commonly used aliases, or spelling variations.

2. Category Consistency: Whether the entity categories are consistent or highly related.

3. Description Relevance: Whether the entity descriptions refer to the same object (e.g., overlapping functions, features, or semantic meaning).

4. Contextual Relationships: Using the provided passage, determine whether the entities refer to the same object in context.

Step 3: Merging. After manual inspection, the entity merging phase begins, updating the entities and relations in the knowledge graph based on the confirmed results. The specific operations are as follows:

- **Entity Name Standardization:** All entity names are standardized to avoid matching issues caused by case differences.
- **De-duplication and Fusion:** The duplicate entities and relations are removed through the merge results, ensuring each entity appears only once in the graph, while updating the descrip-

tion of each merged entity.

- **Knowledge Graph Update:** The merged entities and relations are stored into the respective knowledge graphs, ensuring that the entities and relationships are unique and standardized.

Step 4: Generation. In this step, LLM is used to generate the final alignment results, i.e., the alignment between image entities and text entities. The specific operations are as follows:

- **Image and Text Entity Alignment:** The LLM analyzes the entity information in the images and aligns it with the entities in the text chunks. The matching results for each image entity with the corresponding text entity are generated.
- **Generation of Final Results:** The generated alignment results are saved as `aligned_text_entity.json` files, ensuring that the entity information between the images and text is accurately aligned.

The prompt for fusion is as follows

Prompt for Entity Fusion

-Task-

Merge the text entities extracted from images and the entities extracted from nearby text (chunks). The two sets of entities should be merged based on context, avoiding duplication, and ensuring that each merged entity is derived from both image entities and text entities.

-Explanation-

1. Analyze the entities from the image and the entities from the nearby text, identifying which ones share overlapping or complementary context.

2. Merge entities only if there is a clear contextual link between them (e.g., they describe the same object, concept, or entity). Avoid creating a merged entity if it does not involve contributions from both sources.

3. For each pair of entities that are merged, output the unified entity name, category, the integrated description, and the original sources of the entities involved.

4. Discard entities that cannot be meaningfully merged (i.e., if no matching entity exists in the other source).

-Input Format-

Image Entities:

```
[{"entity_name": "Entity1", "entity_type": "Category1", "description": "Description1", "entity_name": "Entity2", "entity_type": "Category2", "description": "Description2", ...}]
```

Original Text:

[Here is a paragraph of text that provides context for the reasoning.]

Nearby Text Entities:

```
[{"entity_name": "Entity3", "entity_type": "Category3", "description": "Description3", "entity_name": "Entity4", "entity_type": "Category4", "description": "Description4", ...}]
```

-Output Format-

```
[{"entity_name": "Unified Entity Name", "entity_type": "Category", "description": "Integrated Description", "source_image_entities": ["Entity1"], "source_text_entities": ["Entity2"], ...}]
```

The example provided in the prompt is intended to illustrate the task requirements and assist the LLM in understanding the specific objectives. And the example in the prompt is as follows:

```
-Example Input-
Image Entities:
[
{"entity_name": "Electric Sedan", "
  entity_type": "Product", "description":
  "A high-end electric car focusing on
  performance and design"}
]

Original Text:
Tesla has a leading position in the global
electric car market, with its Model S
being a luxury electric vehicle equipped
with advanced autonomous driving
technology and excellent range.

Nearby Text Entities:
[
{"entity_name": "Tesla", "entity_type": "
  Company", "description": "A well-known
  American electric car manufacturer"},
{"entity_name": "Model S", "entity_type": "
  Product", "description": "A luxury
  electric vehicle released by Tesla"}
]

-Example Output-
{"merged_entity_name": "Model S", "
  entity_type": "Product", "description":
  "Model S is a luxury electric vehicle
  released by Tesla, equipped with
  advanced autonomous driving technology
  and excellent range.", "
  source_image_entities": ["Electric Sedan
  "], "source_text_entities": ["Model S"]}
```

Step 5: Check 2. After generating the results, potential hallucination errors generated by the LLM (such as incorrect entity alignments) need to be screened and corrected. The specific operations are as follows:

- **Error Screening:** Check the alignment results generated by the LLM to identify any errors in the fused entity pairs. Ensure that entities requiring fusion actually exist.
- **Random Check:** A random sample comprising 20% of the data is manually reviewed to evaluate both the completeness and accuracy of the entity fusion process. Completeness refers to the proportion of entities that required fusion and were successfully merged, while accuracy pertains to the correctness of the merged entities. The results are shown in Table 5.

Performance	Type			Total.(196)
	News(26)	Aca.(61)	Nov.(109)	
Coverage	86.7	90.0	87.1	90.7
Accuracy	100	99.1	98.4	99.0

Table 5: Manual Inspection Results

In this dataset, we are not concerned with the fusion results of the entities, but rather focus on whether the entities that need to be fused are correctly aligned. Therefore, it can also be said that we are concerned with alignment. As a result, in this paper, we almost do not distinguish between the differences of fusion and alignment.

Complete Results of the CMEL Dataset

In the fusion experiment, we selected different models for testing. The embedding-based similarity methods used three models: all-MiniLM-L6-v2(Yin and Zhang 2024) (MLM), bge-m3(Chen et al. 2024a) (BGE), and stella-en-1.5B-v5(Zhang et al. 2024) (Stella). For the LLMs, we chose Llama3.1-70B-Instruct(AI 2024a) (L) and Qwen2.5-72B-Instruct(Yang et al. 2024) (Q), while for the MLLMs, we selected Qwen2-VL-72B(Wang et al. 2024b) (Qvl) and InternVL2.5-38B-MPO(Chen et al. 2024b) (Intvl).

For the clustering-based approach, the embedding model employed was uniformly Stella-EN-1.5B-V5. After clustering, CMEL requires selecting the appropriate cluster for the target image entity, with two specific methods: KNN(Guo et al. 2003) (K) and LLM-based judgment (L). The LLM utilized was Qwen2.5-72B-Instruct.

The displayed experimental results represent the best outcomes from three runs, and the full results are presented in Table 6.

Method	micro/macro Acc.			Overall.
	News	Aca.	Nov.	
Embedding Model-based Methods				
MLM	2.2/1.7	15.4/14.9	3.9/2.8	9.0/6.5
BGE	6.5/5.7	26.9/26.5	9.3/8.4	17.0/13.5
Stella	10.8/8.4	33.1/34.5	9.0/7.5	20.0/16.8
LLM-based Methods				
L-Qvl	10.8/8.4	33.1/34.5	9.0/7.5	20.0/16.8
L-Intvl	10.8/16.7	30.2/30.0	13.5/13.3	20.8/16.8
Q-Qvl	31.2/24.1	32.2/33.3	19.4/23.2	26.1/26.8
Q-Intvl	33.3/24.1	36.8/36.1	17.4/20.8	27.1/27.0
Clustering-based Methods				
DB-K	48.4/43.1	57.0/58.4	29.9/31.3	43.5/44.3
DB-L	53.8/45.9	60.8/58.3	29.9/34.2	45.2/46.1
KM-K	48.4/41.5	58.2/59.4	29.6/29.4	43.9/43.4
KM-L	50.5/40.6	60.7/57.7	29.6/30.5	45.2/43.0
PR-K	50.5/43.0	61.0/56.4	29.2/33.2	44.7/44.2
PR-L	51.6/44.4	59.7/56.8	29.1/35.2	44.1/45.5
Lei-K	50.5/42.1	66.7/64.3	30.4/37.2	47.7/47.9
Lei-L	54.8/44.7	60.5/55.5	29.4/30.6	44.8/43.6
Spe-K	<u>57.5/50.9</u>	<u>70.1/66.1</u>	<u>31.0/39.8</u>	<u>49.7/55.1</u>
Spe-L	65.5/56.9	73.3/69.9	31.2/39.4	51.8/59.2

Table 6: Complete Results for CMEL dataset

The performance differences across embedding models are quite pronounced: in particular, the relatively weak all-MiniLM-L6-v2 model nearly fails to achieve effective cross-modal entity alignment when using embedding-based similarity methods. The same pattern holds for LLM-based approaches: Llama 3.1-70B-Instruct performs significantly

worse than Qwen 2.5-72B-Instruct in both the news and novel domains, indicating that CMEL task performance is heavily influenced by model architecture and capability. By contrast, clustering-based methods yield more consistent results on the CMEL task. Overall, assigning categories to image entities via LLMs outperforms KNN-based assignment, although the difference is modest. Moreover, the KNN approach requires one fewer model invocation and runs faster, so either method can be chosen flexibly depending on practical considerations and domain requirements.

Complete Results of Multimodal DocQA Experiments

This experiment selects a variety of models as comparison benchmarks, including LLMs such as Llama3.1-70B-Instruct(AI 2024a) (L), Qwen2.5-72B-Instruct(Yang et al. 2024) (Q), and Mistral-Large-Instruct-2411(AI 2024b) (M), as well as MLLMs such as Ovis1.6-Gemma2-27B(Lu et al. 2024) (Ovis), Qwen2-VL-72B(Wang et al. 2024b) (Qvl), and InternVL2.5-38B-MPO(Chen et al. 2024b) (Intvl).

Model	Type				Domain			Overall Acc.
	Aca.	Fin.	Gov.	Laws	News	Text.	Multi.	Una.
LLM-based Methods								
Llama	43.9	13.5	53.4	44.5	<u>79.7</u>	52.9	18.8	81.5
Qwen	41.3	16.3	50.7	49.7	77.3	53.9	20.1	75.8
Mistral	32.3	13.2	43.9	36.1	58.1	43.0	14.6	70.2
MMLLM-based Methods								
Ovis	16.2	11.1	23.6	25.7	39.0	22.8	8.8	54.8
Qvl	17.5	14.9	25.0	34.6	48.8	34.0	8.4	40.3
Intvl	19.8	16.3	28.4	31.4	46.5	35.7	15.9	39.5
NaiveRAG-based Methods								
Llama	43.6	38.2	66.2	64.9	80.2	79.9	32.1	70.2
Qwen	43.6	34.4	62.8	65.4	75.0	<u>81.6</u>	30.5	67.7
Mistral	44.9	35.4	58.1	62.3	76.7	76.5	32.1	69.4
GraphRAG-based Methods								
Llama	40.6	27.1	56.8	59.7	75.0	73.5	24.4	<u>76.6</u>
Qwen	39.6	25.7	52.5	49.7	74.5	71.7	26.0	67.5
Mistral	37.0	28.8	59.2	61.1	75.6	67.7	26.1	76.5
MMGraphRAG-based Methods (Ours)								
L-Ovis	49.7	43.6	58.5	60.0	75.3	75.1	62.5	76.3
Q-Ovis	50.3	46.4	59.4	56.1	76.8	76.3	63.4	74.2
M-Ovis	47.9	40.6	58.6	59.3	75.2	72.6	58.6	74.1
L-Qvl	51.8	59.4	62.8	60.7	77.9	79.1	77.8	70.2
Q-Qvl	51.8	62.9	66.9	<u>68.6</u>	76.2	82.4	81.1	67.7
M-Qvl	48.4	52.8	57.9	62.7	74.5	77.0	75.4	69.6
L-Intvl	60.7	<u>64.1</u>	62.6	64.9	76.2	80.0	<u>86.4</u>	77.5
Q-Intvl	<u>60.5</u>	65.8	<u>66.5</u>	70.4	77.1	81.2	88.7	<u>76.8</u>
M-Intvl	56.4	58.1	58.0	60.2	75.2	76.6	84.9	73.3

Table 7: Complete Results of DocBench Dataset. Based on the experimental results from a total of 21 combinations of the six models, designating Llama3.1-70B-Instruct as the evaluation model does not show undue favoritism towards its own generated results, thereby avoiding erroneous evaluation outcomes.

Among them, Ovis1.6-Gemma2-27B is deployed using the AutoModelForCausalLM from the Transformers(Wolf

et al. 2020) library, InternVL2.5-38B-MPO is deployed using lmdeploy(Chen et al. 2024b), and the other models are deployed using vllm(Kwon et al. 2023).

The complete results for the DocBench dataset are shown in Table 7, the complete results for the MMLongbench dataset are shown in Table 8, and the results for the MM-LongBench dataset by domain are shown in Table 9.

Model	Locations			Modalities				Overall	
	Sin.	Mul.	Una.	Cha.	Tab.	Txt.	Lay.	Fig.	F1
LLM-based Methods									
Llama	23.7	20.3	51.6	16.3	12.7	32.1	21.9	17.4	28.2
Qwen	22.5	20.0	53.2	16.8	12.6	<u>33.3</u>	23.5	16.1	27.8
Mistral	19.1	19.2	38.6	15.6	9.4	28.9	19.2	16.2	23.1
MMLLM-based Methods									
Ovis	10.6	9.5	13.0	6.9	3.1	10.0	15.4	15.4	10.7
Qvl	10.8	9.9	8.1	7.6	5.4	10.0	8.8	12.7	10.0
Intvl	13.3	7.9	13.9	8.8	8.1	10.7	11.8	11.4	11.6
NaiveRAG-based Methods									
Llama	24.9	19.5	56.5	16.3	15.3	31.0	22.7	18.7	29.2
Qwen	22.3	16.4	52.5	15.1	10.8	30.3	20.3	14.9	26.2
Mistral	21.6	19.2	52.9	15.4	12.5	31.2	20.6	13.7	27.3
GraphRAG-based Methods									
Llama	16.3	12.3	<u>78.5</u>	7.6	6.7	25.1	15.0	10.6	27.2
Qwen	18.2	13.2	77.1	14.0	11.0	26.1	12.2	8.5	28.1
Mistral	13.8	10.6	86.5	9.0	5.2	21.7	13.4	7.6	27.2
MMGraphRAG-based Methods (Ours)									
L-Ovis	36.9	13.0	57.4	24.4	23.7	26.8	15.9	24.1	31.0
Q-Ovis	37.4	15.5	54.4	24.9	23.3	28.0	26.1	27.6	31.6
M-Ovis	34.7	12.0	60.0	25.6	21.2	25.7	14.8	22.1	29.5
L-Qvl	37.6	13.8	55.2	26.4	29.2	26.4	13.8	21.2	32.6
Q-Qvl	<u>38.7</u>	20.1	51.6	26.2	30.0	29.3	29.1	<u>29.2</u>	34.8
M-Qvl	36.4	12.1	56.8	27.6	27.7	23.2	11.8	19.8	31.7
L-Intvl	38.7	<u>21.9</u>	59.2	34.7	36.6	31.9	12.9	28.5	<u>36.9</u>
Q-Intvl	39.6	26.7	55.8	<u>34.7</u>	<u>36.5</u>	33.8	28.7	34.6	38.8
M-Intvl	37.7	19.5	63.2	35.0	33.5	28.0	13.6	27.6	35.7

Table 8: Complete Results of MMLongBench Dataset. The GraphRAG-based method has a similar accuracy to the NaiveRAG-based method, but the F1 score is much different. This is because GraphRAG performs better in answering the Una category questions, which are the questions that cannot be answered. Therefore, its overall performance is somewhat worse. The MMGraphRAG-based method significantly outperforms both in accuracy and F1 score, indicating that it can better answer general questions and achieves better overall results.

Through comparative experiments with other single-modal and multi-modal models under various methods, we found that Llama3.1-70B-Instruct is capable of maintaining a degree of independence and objectivity during the evaluation process. This suggests that its evaluation mechanism can effectively distinguish the generation results of different models without bias arising from its own model origin. Therefore, it can be concluded that the evaluation conclusions based on Llama3.1-70B-Instruct are relatively reliable and can provide fair and accurate assessment results in multi-model document QA experiments.

Model	Evidence Locations							Overall	
	Int.	Tut.	Aca.	Gui.	Bro.	Adm.	Fin.	Acc.	F1
LLM-based Methods									
Llama	33.5	32.1	27.7	24.7	22.0	31.1	18.8	28.2	23.0
Qwen	31.5	31.8	25.4	30.3	26.5	36.6	9.6	27.8	22.1
Mistral	28.2	25.7	19.3	22.0	22.9	32.8	8.4	23.1	19.2
MMLLM-based Methods									
Ovis	7.4	20.4	10.9	7.3	16.5	14.3	4.3	10.7	9.2
Qvl	7.8	20.1	6.8	10.6	9.8	11.6	7.2	10.0	9.5
Intvl	11.3	19.1	10.4	8.9	14.4	13.3	6.0	11.6	10.4
NaiveRAG-based Methods									
Llama	34.0	31.0	30.0	29.4	23.0	29.6	18.4	29.2	24.2
Qwen	30.0	27.6	25.9	31.1	17.0	32.1	12.8	26.2	20.9
Mistral	32.6	24.7	24.5	32.5	18.7	34.0	17.6	27.3	22.8
GraphRAG-based Methods									
Llama	30.8	27.0	25.0	29.7	24.0	34.4	16.7	27.2	18.2
Qwen	34.6	21.8	24.8	29.2	27.0	<u>36.2</u>	18.9	28.1	19.3
Mistral	34.4	22.3	24.9	28.1	26.0	31.6	15.5	27.2	16.8
MMGraphRAG-based Methods (Ours)									
L-Ovis	35.7	32.9	30.2	<u>41.7</u>	27.1	24.6	26.7	31.0	26.8
Q-Ovis	36.4	38.9	29.4	39.3	<u>32.2</u>	30.2	30.1	31.6	27.5
M-Ovis	36.5	34.2	29.4	38.3	31.5	22.5	25.5	29.5	25.5
L-Qvl	36.5	32.9	27.1	41.3	26.3	28.4	34.0	32.6	28.1
Q-Qvl	37.2	38.8	26.3	39.5	31.6	35.8	<u>38.0</u>	34.8	30.4
M-Qvl	37.0	35.2	26.5	38.8	30.9	26.4	32.2	31.7	26.9
L-Intvl	42.4	34.0	36.8	42.8	28.9	26.0	36.2	<u>36.9</u>	<u>32.4</u>
Q-Intvl	43.3	36.5	35.2	40.0	32.7	33.5	41.0	38.8	34.1
M-Intvl	<u>42.9</u>	35.0	<u>35.6</u>	38.9	31.1	23.2	35.2	35.7	31.8

Table 9: MMLongBench Dataset Domain Results. The MM-GraphRAG method achieves the best performance across all domains except for the Administration/Industry file category. Notably, it demonstrates the most significant improvements in the Guidebook and Financial report categories, which are characterized by a high volume of charts, tables and figures. These enhancements are far more pronounced than those of other methods.

The MMLongBench dataset encompasses a diverse range of document domains. These domains include Research Reports/Introductions (Int.), which typically feature academic or industry-oriented analyses and background information; Tutorials/Workshops (Tut.), focusing on instructional content for skill development or knowledge dissemination; Academic Papers (Aca.), containing scholarly research and findings; Guidebooks (Gui.), offering practical information and advice for specific topics or activities; Brochures (Bro.), designed for promotional or informational purposes in a concise format; Administration/Industry Files (Adm.), covering official documents or industry-specific reports; and Financial Reports (Fin.), presenting financial data and analyses.

Generation Module

Given that the current technology for MLLMs is still not fully mature, the Generation module employs a hybrid-generation approach that combines results from an LLM (text-only model) and an MLLM (capable of processing text, images, and other multi-modal information). This strategy

aims to enhance the accuracy and stability of the final output.

This hybrid strategy not only makes full use of the maturity and advantages of LLMs in specific tasks, ensuring high quality semantic understanding and language expression, but also enriches the generated content by integrating multi-modal information. This approach effectively mitigates the limitations of current MLLMs while ensuring an overall improvement in the quality of the system’s generation.

The Generation module achieves hybrid-generation through five steps: First, it extracts relevant information from the output of retrieval module and generates an initial textual response using LLM. Next, the MLLM is invoked to combine image and text information to generate a multi-modal response, which is then merged with other multi-modal responses through an LLM. Finally, the generated textual response and multimodal response are integrated to output a comprehensive and consistent final result.

A specific example of an execution process is as follows, with the user’s query and the generated answer provided:

question: "What is the highest F1 score achieved on the Chinese OntoNotes4.0 dataset, according to Table 10?"

answer: "The highest F1 score on the Chinese OntoNotes4.0 dataset is 84.67."

type: "multi-modal-t"

evidence: "The question requires identifying the maximum value in a single column, which is straightforward."

LLM Response:

The information cannot be obtained from the provided data tables.

MLLM Responses:

response 1

*To answer the question about the highest F1 score achieved on the Chinese OntoNotes 4.0 dataset, we will refer to the table labeled "Table 10" in the image provided. This table displays the effects of varying hyperparameters, specifically the alpha values, on the performance metrics of two datasets: Chinese OntoNotes 4.0 and English QuoRef. The table columns include alpha values ranging from 0.1 to 0.9 in increments of 0.1. For each alpha value, there are corresponding F1 scores for both Chinese OntoNotes 4.0 and English QuoRef datasets. Upon examining the "Chinese OntoNotes 4.0" column in the table, we notice the following F1 scores for each alpha value:- alpha = 0.1: F1 score = 80.13 - alpha = 0.2: F1 score = 81.17 - alpha = 0.3: F1 score = 84.22 - alpha = 0.4: F1 score = 84.52 - alpha = 0.5: F1 score = 84.47 - alpha = 0.6: F1 score = 84.67 (bolded, indicating the highest score) - alpha = 0.7: F1 score = 81.81 - alpha = 0.8: F1 score = 80.97 - alpha = 0.9: F1 score = 80.21 From this data, it is clear that the highest F1 score for the Chinese OntoNotes 4.0 dataset is ****84.67****, which is achieved at an alpha value of 0.6. This score is highlighted in bold in the table, further confirming its significance as the peak performance metric for this specific dataset within the given range of alpha values.*

response 2

According to the table in image 12, the highest F1 score achieved on the Chinese OntoNotes 4.0 dataset is 84.67, which occurs at an alpha value of 0.6. This table illustrates the effects of hyperparameters on the performance metrics of the Chinese OntoNotes 4.0 and English QuoRef datasets, showing that the F1 score for Chinese OntoNotes 4.0 peaks at 84.67 when (alpha) is set to 0.6.

response 3

I'm sorry, but I cannot provide an analysis of the table in the image as it was not provided.

Merged MLLM Response:

The highest F1 score achieved on the Chinese OntoNotes 4.0 dataset is 84.67, which occurs at an alpha value of 0.6. This information is confirmed by the data in Table 10, which displays the effects of varying hyperparameters, specifically the alpha values, on the performance metrics of two datasets: Chinese OntoNotes 4.0 and English QuoRef.

Final Result:

The highest F1 score achieved on the Chinese OntoNotes 4.0 dataset, according to analysis of the information provided, is 84.67. This optimal F1 score occurs when the alpha value is set to 0.6.

The prompt for merging multimodal responses is as follows:

Prompt for Merging MLLM responses

The following is a list of responses generated by a multi-modal model based on the same user Query but different images. Please perform the following tasks:

- Analyze the Responses: Identify any contradictions, repetitions, or inconsistencies among the responses.

- Reasonably Determine: Decide which response best aligns with the user Query based on the provided information, ensuring that the determination is based on the relevance and accuracy of the information in the response rather than a majority consensus, as the correct answer may only pertain to a specific image and may not align with the majority.

- Provide a Unified Answer: Deliver a single, unified response that eliminates contradictions, resolves ambiguities, and accurately addresses the user Query.

Additionally, retain any highly relevant information from the responses that supports or complements the unified answer.

MMKG Construction Paradigms: N-MMKG vs. A-MMKG

In traditional multimodal knowledge graph (MMKG) construction, visual and textual information are typically treated as distinct modalities and incorporated into a unified graph structure in the form of entity attributes. However, in the context of complex cross-modal reasoning and information fusion, such attribute-centric MMKGs (A-MMKG) often

lead to semantic loss—particularly when modeling intricate relationships between images and text (Liang et al. 2024). To address this limitation, we adopt a node-based MMKG (N-MMKG) paradigm as the foundation of our graph construction.

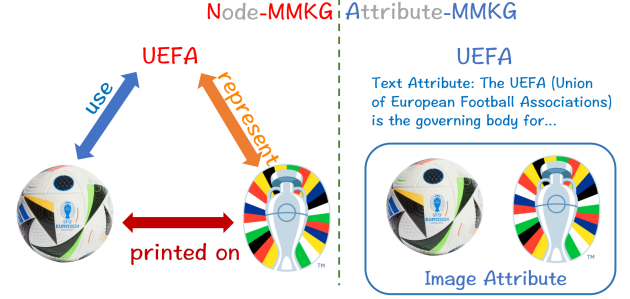


Figure 6: N-MMKG vs. A-MMKG

In N-MMKG, images are treated as standalone nodes, thereby avoiding the information loss associated with storing visual data merely as attributes of textual entities. For example, a visual input related to the "UEFA" can be represented as an independent node, which enables explicit modeling of the "used in" relation between "football" and "UEFA", as well as the "symbolizes" relation between a "logo" and "UEFA". This representation preserves richer semantic information. Furthermore, this design allows an image to be simultaneously associated with multiple textual entities and easily integrated with other modalities (e.g., audio or video). New modalities can be seamlessly added as independent nodes without requiring structural modifications to the existing graph, significantly enhancing the flexibility and scalability of MMKG.

Implementation of the Fusion Module

Algorithm 1: Fusion

```

1: function FUSION(images)
2:   Initialize  $imgdata \leftarrow$  image_data.json
3:   Initialize  $tkg \leftarrow$  chunk_knowledge_graph.json
4:   Initialize  $chunks \leftarrow$  text_chunks.json
5:   for each  $image \in$  images do
6:     if  $mmkg$  exists then
7:       continue to next iteration
8:     end if
9:      $ikg \leftarrow$  FIND( $imgdata, image$ )
10:     $list^a \leftarrow$  ALIGN( $tkg, ikg$ )
11:     $ikg_e \leftarrow$  ENHANCE( $list^a, ikg, tkg, chunks$ )
12:     $ikg_u \leftarrow$  UPDATE( $ikg_e, tkg, image, chunks$ )
13:     $mmkg \leftarrow$  MERGE( $ikg_u, tkg, list^a$ )
14:  end for
15:   $vdb \leftarrow mmkg$ 
16:  return  $mmkg, vdb$ 
17: end function

```

The pseudocode for the entire fusion process is as above, which provides a clearer understanding of the input and out-

put of each step. The final step utilizes the fused MMKG to construct an entity vector database (vdb), which facilitates the retrieval stage.

imgdata (image data) and *chunks* (text chunks) are both results of preprocessing, stored in their respective JSON files. *imgdata* is used to store various information related to images, while *chunks* store the text chunks of the entire document. *tkg* (text-based knowledge graph) is the result of the text modality processing module *txt2graph*. It is stored in JSON files for each chunk and also as a complete GraphML file. Here, we do not make a specific distinction between them. *ikg* represents the image-based knowledge graph. First, *ikg* is obtained from the *image* and *imgdata*. Then, the formal first step is to align the entities in *tkg* and *ikg*, and save the alignment results as *list_a*. Entities to be fused are filtered out from *ikg*, and the remaining entities are enhanced using *chunks* (context) and relevant entities from *tkg*, resulting in the enhanced image-based knowledge graph *ikg_e*.

Next, we perform a detailed search in *tkg* to align the global entity of the image. This involves extracting relevant text segments from chunks that are semantically related to the whole image. Once the entity is identified, matching is performed in *tkg*. If aligned entity is found, only the relations of *ikg_e* will be updated to obtain *ikg_u*, which naturally achieves alignment during fusion. If no aligned entity is found in *tkg*, *ikg_e* will be updated by supplementing a new entity obtained from chunks to form *ikg_u*. Finally, *tkg* and *ikg_u* are fused based on the results of *list_a* to obtain the final *mmkg* (multimodal knowledge graph).

The second step, entity enhancement, is achieved using the reasoning capabilities of LLMs. Based on the context information, entities from *tkg* are used to supplement visual entities that do not have aligned counterparts. The prompt to enhance entities is as follows:

Prompt for Enhancing Image Entities

The goal is to enrich and expand the knowledge of the image entities listed in the *img_entity_list* based on the provided *chunk_text*.

The *entity_type* should remain unchanged, but you may modify the *entity_name* and *description* fields to provide more context and details based on the information in the *chunk_text*.

For each entry in the *img_entity_list*, the following actions should be performed:

1. Modify and enhance the *entity_name* if necessary.
2. Expand the *description* by integrating relevant details and insights from the *chunk_text*.
3. Include an *original_name* field to capture the original entity name before enhancement.

Ensure the final output is in valid JSON format, only including the list of enhanced entities without any additional text.

After generating candidate entities, LLM is used to align visual entities, with the specific prompt as follows:

Prompt for Generating Image Feature Block Description

You are an expert system designed to identify matching entities based on semantic similarity and context. Given the following inputs:

img_entity: The name of the image entity to be evaluated.

img_entity_description: A description of the image entity.

chunk_text: Text surrounding the image entity providing additional context.

possible_image_matched_entities: A list of possible matching entities. Each entity is represented as a dictionary with the following fields:

entity_name: The name of the possible entity.

entity_type: The type/category of the entity.

description: A detailed description of the entity.

additional_info: Additional relevant information about why choose this entity (such as similarity, reason generated by LLM, etc.).

-Task-

Using the information provided, determine whether the *img_entity* matches any of the entities in *possible_image_matched_entities*. Consider the following criteria:

1. Semantic Matching: Evaluate the semantic alignment between the *img_entity* and the possible matching entities, based on their names, descriptions, and types. Even without a similarity score, assess how well the *img_entity* matches the attributes of each possible entity.

2. Contextual Relevance: Use the *chunk_text* and *img_entity_description* to assess the contextual alignment between the *img_entity* and the possible entity.

-Output-

If a match is found, only return the *entity_name* of the best-matching entity.

If no match meets the criteria (e.g., low similarity or poor contextual fit), only output "no match".

Do not include any explanations, reasons, or additional information in the output.

Implementation of the Img2Graph Module

Image Segmentation. The initial step involves subjecting the input image to segmentation, which entails parsing the image into regions that possess distinct semantic meanings (referred to as image feature blocks). This process is achieved through the image segmentation function in the YOLOv8 (Jocher, Chaurasia, and Qiu 2023) model. The segmentation outcomes provide a finer-grained representation of image information, thereby establishing the foundation for subsequent entity extraction and relation modeling. The granularity of segmentation significantly influences the precision of edge delineation within knowledge graphs.

Image Feature Block Description. Subsequently, the MLLM is employed to generate text descriptions for each image feature block derived from the segmentation results. The generation process of the model is guided by carefully designed prompts, which enable it to produce accurate and semantically rich information. These text descriptions not

only construct independent entities for the image modality but also provide a bridge for subsequent alignment with the text modality. Specific prompts are as follows:

Prompt for Generating Image Feature Block Description

You are a multi-modal model capable of processing image feature blocks and generating detailed descriptions.

Your task is to first determine the category of the given image feature block (object, organism, or person) and then extract the entity's features from it, providing a detailed description.

Note that the entity in the image feature block may not be complete, such as a half-body photo of a person or a partial image of an object.

- If the entity is an object, describe the object's features, including its name, color, shape, size, material, possible function, and other significant characteristics.

- If the entity is an organism, describe the features of this organism (animal or plant), including species, name, age, color, shape, size, posture, or structural characteristics.

- If the entity is a person, describe the person's features, including gender, skin color, hairstyle, clothing, facial expression, age, and posture.

All image feature blocks have a black background, so focus solely on the entity's characteristics, and do not mention "the background is black" in the output.

The example in the prompt is derived from the processing results illustrated in Methodology, and the example is as follows:

```
The category of this image feature block is
'person'. The entity features are as
follows:
Person Features:
- Gender: Female
- Hairstyle: Long hair, light brown,
  naturally falling with some hair pinned
  with a clip
- Eyes: Blue, large and expressive
- Expression: Smiling, appears friendly and
  joyful
- Age: Appears to be a young woman
- Clothing: Wearing a white shirt with the
  sleeves rolled up, revealing the wrists;
  paired with blue overalls, with dark
  blue straps; light blue sneakers on her
  feet
- Accessories: Orange shoulder bag on her
  right shoulder; brown belt tied around
  the waist
- Holding: Holding a vintage-style camera
  with both hands, the camera is black and
  silver, with a large lens, appearing
  professional
Overall, the character gives off a youthful,
lively vibe with a touch of artistic
flair.
```

Entity and Relation Extraction from the Image. This step employs an MLLM, guided by prompts, to identify explicit relations (e.g., "girl — girl holding a camera — camera") and implicit relations (e.g., "boy — the boy and girl appear to be close, possibly friends or a couple — girl"). The extracted entities and relations provide structured information for the multimodal extension of the knowledge graph.

Compared to traditional scene graph generation methods, MLLM-based approaches excel at extracting entities and inferring both explicit and implicit relations by leveraging the semantic segmentation and reasoning abilities of MLLMs, resulting in high-precision and fine-grained scene graphs.

The prompt for extracting entity and relation from image is as follows:

Prompt for Visual Entity and Relation Extraction

Given a raw image, extract the entities from the image and generate detailed descriptions of these entities, while also identifying the relationships between the entities and generating descriptions of these relationships. Finally, output the result in a standardized JSON format. Note that the output should be in English.

-Steps-

1. Extract all entities from the image.

For each identified entity, extract the following information:

- Entity Name: The name of the entity

- Entity Type: Can be one of the following types:

[[entity.types]]

- Entity Description: A comprehensive description of the entity's attributes and actions

- Format each entity as ("entity"{tuple_delimiter}

Entity_Name {tuple_delimiter} Entity_Type {tuple_delimiter} Entity_Description

2. From the entities identified in Step 1, identify all pairs of (Source Entity, Target Entity) where the entities are clearly related.

For each related pair of entities, extract the following information:

- Source Entity: The name of the source entity, as identified in Step 1

- Target Entity: The name of the target entity, as identified in Step 1

- Relationship Description: Explain why the source entity and target entity are related

- Relationship Strength: A numerical score indicating the strength of the relationship between the source and target entities

Format each relationship as ("relationship" {tuple_delimiter} Source_Entity {tuple_delimiter} Target_Entity {tuple_delimiter} Relationship_Description {tuple_delimiter} Relationship_Strength)

3. Return the output as a list including all entities and relationships identified in Steps 1 and 2. Use {record_delimiter} as the list separator.

4. Upon completion, output {completion_delimiter}

The examples contained within the prompt are excessively lengthy. For illustrative purposes, only a small excerpt is presented here to demonstrate the format, as follows:


```

("entity"{tuple_delimiter}"Girl"{
  tuple_delimiter}"person"{tuple_delimiter}
  "Wearing glasses, dressed in black,
  holding white and blue objects, smiling
  at the camera."){record_delimiter}
("entity"{tuple_delimiter}"Headphones"{
  tuple_delimiter}"object"{tuple_delimiter}
  "White headphones on the girl's ears.")
{record_delimiter}
...
("relationship"{tuple_delimiter}"Girl"{
  tuple_delimiter}"Headphones"{
  tuple_delimiter}"The girl is wearing
  headphones."{tuple_delimiter}8){
  record_delimiter}
("relationship"{tuple_delimiter}"Girl"{
  tuple_delimiter}"Phone"{tuple_delimiter}
  "The girl is holding a phone in her hand
  ."{tuple_delimiter}8){record_delimiter}
...

```

Alignment of Image Feature Blocks with Entities. Based on the extracted visual entities, the feature blocks generated by segmentation are aligned with their corresponding textual entities. This step is accomplished through the recognition and reasoning capabilities of the MLLM. For example, based on the semantic content of the textual entity, "Feature Block 2" is identified as the image of a "boy," and a relation is established in the knowledge graph. This alignment not only connects feature blocks to entities but also strengthens the association between modalities.

The prompt for aligning visual entities is as follows:

Prompt for Visual Entity Alignment

-Objective-

Given an image feature block and its name placeholder, along with entity-description pairs extracted from the original image, determine which entity the image feature block corresponds to and output the relationship with the entity. The output should be in English.

-Steps-

1. Based on the provided entity-description pairs, determine the entity corresponding to the image feature block and output the following information:

- Entity Name: The name of the entity corresponding to the image feature block

2. Output the relationship between the image feature block and the corresponding entity, and extract the following information:

- Image Feature Block Name: The name of the input image feature block

- Relationship Description: Describe the relationship between the entity and the image feature block, with the format "The image feature block Image Feature Block Name is a picture of Entity Name."

- Relationship Strength: A numerical score representing the strength of the relationship between the image feature block and the corresponding entity

Be sure to include the {record_delimiter} to signify the end of the relationship.

The examples saved in the prompt are as follows:

Example 1:

The image feature block is as shown above, and its name is "image_0_apple-0.jpg."

Entity-Description:

"Apple" - "A green apple, smooth surface, with a small stem."

"Book" - "Three stacked books, red cover, yellow inner pages."

Output:

```

("relationship"{tuple_delimiter}"Apple"{
  tuple_delimiter}"image_0_apple-0.jpg"{
  tuple_delimiter}"The image feature block
  image_0_apple-0.jpg is a picture of an
  apple."{tuple_delimiter}7){
  record_delimiter}

```

Global Entity Construction. Finally, a global entity is constructed for the entire image, serving as a global node in the knowledge graph. This node not only provides supplementary descriptions of the image's global information (e.g., "meet on the bridge") but also enhances the completeness of the knowledge graph through its connections to local entities. Through this step, the knowledge graph can provide multi-level information from global to local, further enhancing retrieval capabilities.