MMGraphRAG: Bridging Vision and Language with Interpretable Multimodal Knowledge Graphs

Xueyao Wan
1185909349@qq.com

Hang Yu
hang023@e.ntu.edu.sg

Abstract

Large Language Models (LLMs) excel at natural language generation but often suffer from hallucinations due to their static parametric nature. Retrieval-Augmented Generation (RAG) mitigates this limitation by incorporating external knowledge, while GraphRAG further enhances reasoning by leveraging knowledge graphs (KGs).

One limitation, however, is that methods based on GraphRAG are still limited to textual data because of the difficulty of constructing a fine-grained multimodal KG (MMKG). All existing text–image fusion methods, including mapping images and texts into a shared embedding space, image captioning, and joint text–image extraction, rely heavily on task-specific training and suffer from limited generalization. More importantly, they fail to preserve the structural knowledge of visual content or capture the structure of knowledge and reasoning paths between modalities. Thus, they cannot construct an MMKG that incorporates both image and textual entities.

To address these limitations, we propose MMGraphRAG, which refines visual content through scene graphs and integrates it with text-based knowledge graphs to construct an MMKG via a novel cross-modal fusion approach. It employs the SpecLink method, which leverages spectral clustering-based candidate entities generation method to achieve accurate cross-modal entity linking and retrieves context along reasoning paths to guide the generative process. Due to the lack of benchmarks, we also release the CMEL dataset, which is designed for fine-grained multi-entity alignment in complex multimodal scenarios. The experiments on this dataset validate the effectiveness of the SpecLink method in complex multimodal scenarios. To evaluate MMGraphRAG, we conduct experiments on the DocBench and MMLongBench datasets. Results show that MMGraphRAG achieves state-of-the-art performance, demonstrating strong domain adaptability and multimodal information processing capability.

CCS Concepts

• **Computing methodologies** → **Information extraction**; **Information extraction**; • **Information systems** → *Content analysis and feature selection*; *Document structure*.

Keywords

GraphRAG, Cross-Modal Entity Linking, Scene Graph

1 Introduction

Large Language Models (LLMs) have advanced in natural language generation, yet hallucination—factual inconsistency—remains a major limitation[21, 32, 41]. Due to their static parametric nature, LLMs cannot promptly integrate specialized and up-to-date data, causing incomplete or inaccurate knowledge in professional fields[9, 12, 20, 57]. 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[26]. To improve the performance on complex QA tasks, researchers have integrated knowledge graphs (KGs) to enhance the reasoning capabilities and interpretability of RAG systems, such as GraphRAG, which integrates local and global knowledge by constructing entity KGs and community summaries[15]. However, methods based on GraphRAG are still limited to textual data because of the difficulty of constructing proper multimodal KG (MMKG). Text-only GraphRAG methods cannot fully exploit visual information, leading to incomplete retrieval results[29], since real-world information often co-exists in multimodal forms such as text, images, and tables.

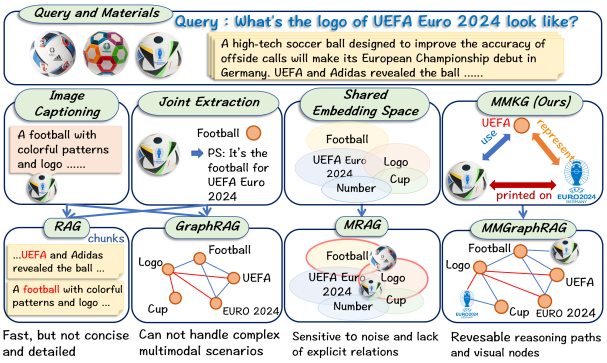

Figure 1: Comparison of Image-Text Fusion Methods. Prior methods struggle with accurate visual reasoning: (a) Captioning-based methods linearize the image into a single text description, irretrievably losing fine-grained details. (b) Joint Extraction depends on precise annotations and fails in complex scenarios where key visual entities, like the logo itself, are unlabeled. (c) Shared Embedding MRAG struggles to isolate specific attributes from a flattened, implicit vector. In contrast, (d) our MMGraphRAG constructs a Multimodal Knowledge Graph, representing the logo and football as explicit visual nodes linked to text. This preserves the essential knowledge structure, enabling precise and interpretable answers.

To bridge this gap, we propose MMGraphRAG framework with a creative and zero-shot construction method of fine-grained MMKG. To facilitate the MMGraphRAG system, the MMKG should construct KG for the image modality and enable fusion connections across image–text modalities which remains graph structure. Our method achieves these by utilizing scene graphs and designing a novel cross modal entity linking (CMEL) method called SpecLink, ensuring that image information is preserved in the MMKG as entities and integrated with textual entities via high-quality links. All

existing alternative multimodal fusion approaches, such as mapping images and text into a shared embedding space[11, 16, 31, 62], image captioning[52, 65], and image-text joint extraction[50, 67], fail to realize these key points and require large amounts of domain-specific training. Figure 1 presents a comparison of these multimodal fusion methods. As shown, the successful construction of the MMKG retains knowledge structure and reasoning paths, allowing the MMGraphRAG approach to deliver more comprehensive and accurate results, thereby mitigating hallucinations.

The key point of constructing MMKG is CMEL task, which enables the establishment of connections between entities from different modalities[60]. 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 introduce the SpecLink method, which leverages spectral clustering to generate high-quality 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.

Evaluated on two challenging benchmarks, MMGraphRAG demonstrates significant superiority. On DocBench[68], it achieves 76.8% overall accuracy, far exceeding NaiveRAG (59.5%) and GraphRAG (52.3%). This advantage is especially pronounced in multimodal queries (88.7% accuracy), showcasing effective visual-textual integration. Similarly, on MMLongBench[39], MMGraphRAG establishes a new state of the art with 38.8% accuracy and a 34.1% F1 score, excelling in complex chart and figure reasoning. Crucially, our framework demonstrates exceptional robustness in identifying unanswerable questions, outperforming a leading multimodal RAG method by over sixfold (35.1% vs. 5.8%). This highlights the effectiveness of its explicit MMKG reasoning in mitigating hallucinations. Further ablation studies confirm that our spectral clustering-based fusion mechanism is critical, preventing the introduction of noise and substantially boosting performance over naive fusion methods. Overall, MMGraphRAG provides not only superior accuracy but also more reliable and interpretable reasoning paths.

Our main contributions are as follows:

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

2 Related Work

2.1 Image-Text Fusion Methods

Research on multimodal fusion as the upstream of RAG has mainly progressed along following directions. First, image captioning methods adopt encoder–decoder frameworks that translate images into textual descriptions, thereby "linearizing" visual semantics before retrieval or reasoning[52]. For example, while recent models incorporate scene graph encoding to better encode object relationships[65], the structural information remains only implicitly folded into the captioning model. The final output is still a free-form textual description, which discards the explicit knowledge structure and lack transparent reasoning paths.

Second, shared embedding space approaches align images and text in a joint vector space using large-scale contrastive pretraining, achieving robust zero-shot retrieval but sensitive to noise and also leading to semantic flattening[11, 16, 31, 62]. The specialization caused by training, along with noisy evidence in a shared embedding space, can mislead the model into confidently generating incorrect answers[40]. For instance, M3DocRAG achieves only 5.8% accuracy on questions with "unanswerable" ground truths, underscoring significant weaknesses in factual consistency [11]. Thus, the limitations of this method in structured reasoning, generalizability, and hallucination reduction call for further improvement.

Third, multimodal alignment with constraints leverages semantic priors to disambiguate entity linking across modalities, such as enforcing consistency between object regions and textual mentions[50], or constraining alignment through syntactic or semantic cues[67]. While such methods improve disambiguation, they rely on task-specific annotations or heuristics and thus suffer from limited generalizability. In contrast, our method explicitly integrates scene graphs into an MMKG by treating visual nodes as first-class entities and linking them to textual counterparts via CMEL. This enables structure-preserving retrieval and interpretable multimodal reasoning while directly utilizing visual information without semantic flattening.

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

In the construction of Multimodal Knowledge Graphs (MMKGs), information can be integrated using two primary paradigms. The conventional attribute-centric MMKG (A-MMKG) approach treats visual and textual information as distinct modalities, where non-textual data is typically embedded as an entity's attribute. However, this method can lead to significant semantic loss, particularly when modeling intricate cross-modal relationships [27]. To address this limitation, our framework adopts a node-based MMKG (N-MMKG) paradigm as its foundation.

The distinction is best illustrated with a medical example. In an A-MMKG, a "Patient" entity might possess an attribute such as "has_scan: mri_001.jpg". This structure relegates the image to a passive descriptor, making it impossible to model a direct relationship from the MRI scan to a specific clinical finding. In contrast, the N-MMKG paradigm elevates the MRI scan to a standalone node. This allows for the explicit modeling of crucial, fine-grained relations, such as "mri_001.jpg shows evidence of Anomaly". The image is no longer just a data point but an active component of the graph's relational structure.

This node-based architecture is fundamentally better suited for our framework because it preserves the rich, cross-modal semantics required for complex reasoning. It also significantly enhances

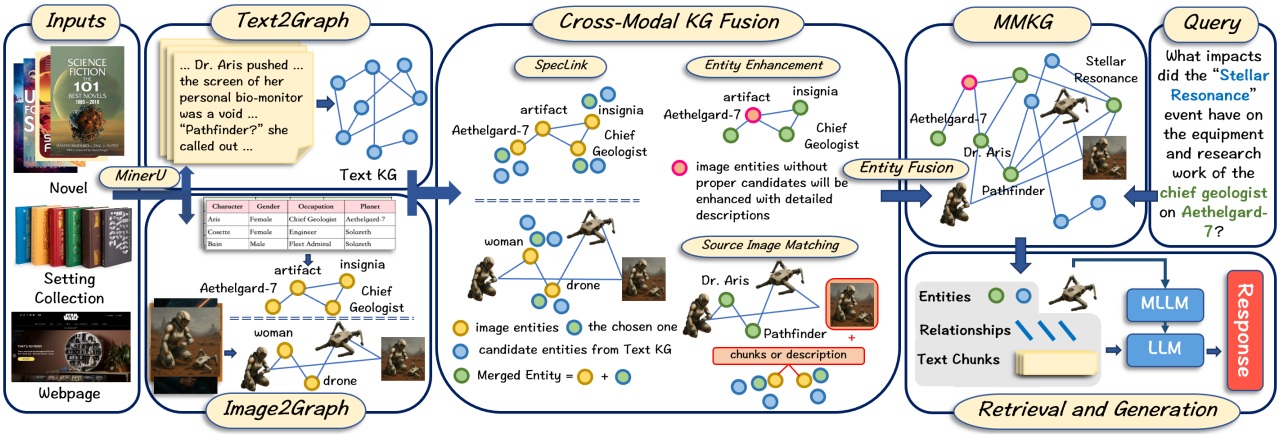

Figure 2: MMGraphRAG Framework Overview. The framework begins by parsing sources like novels and webpages, creating parallel knowledge graphs through Text2Graph for text and Image2Graph for visual content. The central Cross-Modal KG Fusion module then unifies these graphs by intelligently linking entities; for example, it uses SpecLink method to align the textual mention of the "Dr. Aris" with the corresponding "woman" in the image. This process yields a cohesive MMKG. Finally, during the Retrieval and Generation stage, this MMKG provides structured, cross-modal context to an MLLM/LLM system, enabling it to answer complex queries with accurate, well-supported responses.

flexibility and scalability. By design, new modalities can be seamlessly integrated as independent nodes without requiring structural modifications to the existing graph, ensuring a more robust and extensible knowledge foundation.

2.3 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 [42, 43]. MEL extends EL by incorporating visual information as auxiliary attributes to enhance alignment between entities and knowledge base entries [17, 35, 45]. 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 [60]. 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 [5]. To bridge this gap, we construct a CMEL dataset featuring greater real-world complexity and propose SpecLink method to drive further advances in CMEL research.

3 Methodology

To facilitate effective cross-modal information fusion and reasoning, we propose MMGraphRAG, a novel 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.

3.1 MMGraphRAG Framework

The overall architecture of the proposed MMGraphRAG framework is illustrated in Figure 2. It consists of three stages: Indexing, Retrieval, and Generation. In contrast to query-based methods, where a temporary graph is constructed during the retrieval phase based on the query information and the database[6], the MMGraphRAG framework provides a more scalable solution. The MMKG built by MMGraphRAG can function as an independent knowledge base. Once created, it can be used across various queries without requiring re-building, making it ideal for large-scale question-answering scenarios.

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* [53], 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 SpecLink 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.

The retrieval stage extends the traditional GraphRAG by incorporating image-based information, enhancing the retrieved entities

and relations with visual data. In the generation stage, a hybrid approach is used, where the LLM handles the logical flow and textual response generation, while the MLLM focuses on processing the image retrieval results. This approach improves upon the GraphRAG method by adding image information to the retrieval process, allowing for more comprehensive and accurate multimodal response generation.

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. Finally, by constructing the MMKG based on LLMs, the framework eliminates reliance on training, offering enhanced flexibility.

3.2 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 [33, 34].

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 [7, 54]. 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 [30, 46, 47, 58, 63, 66], MLLM-based methods improve generalization capabilities by minimizing human supervision.

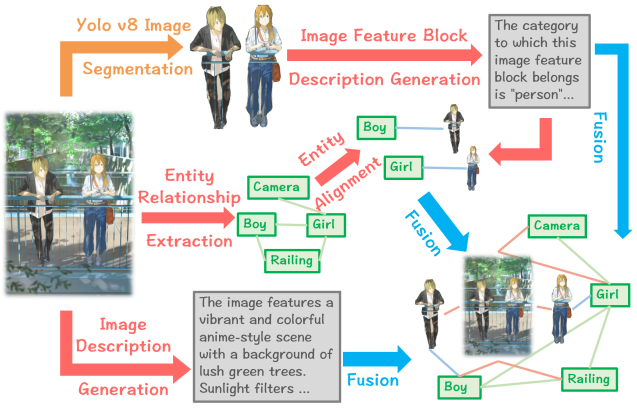

Figure 3: An Example of the Img2Graph Module in Action

The Img2Graph module maps images into KGs through a five-step pipeline. First, semantic segmentation is performed using YOLO [51] 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 3.

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.

3.3 Cross-modal Fusion Module

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:

3.3.1 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. Define image set as $I = \{I_1, I_2, \ldots, I_N\}$, where each image I_i contains a set of extracted entities denoted by: $E(I_i) = \{e_1^{(I_i)}, e_2^{(I_i)}, \ldots, e_K^{(I_i)}\}$.

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

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 $e_k^{(I_i)}$, 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 module to ensure they share a unified representation in the KG.

SpecLink:A Spectral Clustering–Based Candidate Generation. To improve the efficiency and robustness of candidate entities generation, we propose SpecLink, a spectral clustering-based optimization strategy. Existing methods fall into two categories: (1) *distance-based clustering*, such as KMeans [24, 28, 44] and DBSCAN [13, 23], which depends semantic similarity but ignores graph structure, and (2) *graph-based clustering*, such as PageRank [1] and Leiden [49], which captures structural relations but suffers in sparse graphs. To address both aspects, we design SpecLink 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:

$$\mathbf{A}_{pq} = sim(\mathbf{v}_p, \mathbf{v}_q) \cdot weight(r_{pq}) \tag{1}$$

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

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

$$\mathbf{D}_{pp} = \sum_q \mathbf{A}_{pq}.$$

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 $\mathbf{Q} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ using the smallest m eigenvectors, where m depends on the number of textual entities in context [22].

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

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

For each image entity $e_k^{(I_i)}$, we select the most relevant cluster based on the cosine similarity between its embedding $\mathbf{v}_k^{(I_i)}$ 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 [35, 36]. 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.

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

For instance, while the original text may only state that *"HURRICANE SLAMS COASTAL REGIONS IN WEST FLORIDA..."*, the corresponding image depicts *a flooded neighborhood* without explicit textual alignment. By integrating contextual cues such as *Hurricane Ian* and *Florida* from the text, the visual entity can be enriched into *"A neighborhood in Florida severely flooded by Hurricane Ian, causing significant damage and displacement."* This step ensures high-fidelity entity descriptions and enhances the internal completeness of the MMKG.

3.3.3 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.

3.3.4 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.

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

4 Experiments

In this section, we present a comprehensive empirical evaluation of the proposed MMGraphRAG framework. Our evaluation is structured in three main parts. First, we conduct a targeted experiment on the CMEL task to specifically validate the effectiveness of our SpecLink method for generating candidate entities. Second, we assess the end-to-end performance of the complete MMGraphRAG system on two challenging multimodal Document Question Answering (DocQA) benchmarks, DocBench and MMLongBench, comparing it against a suite of baseline methods. Finally, to provide deeper insight into our model's architecture, we conclude with a series of ablation studies that analyze the crucial roles of the cross-modal fusion module and the explicit graph-based representation.

4.1 CMEL Experiments

To evaluate the effectiveness of our SpecLink 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[5]. 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. Appendix provides details on the construction process, dataset description, illustrative example, and reliability validation.

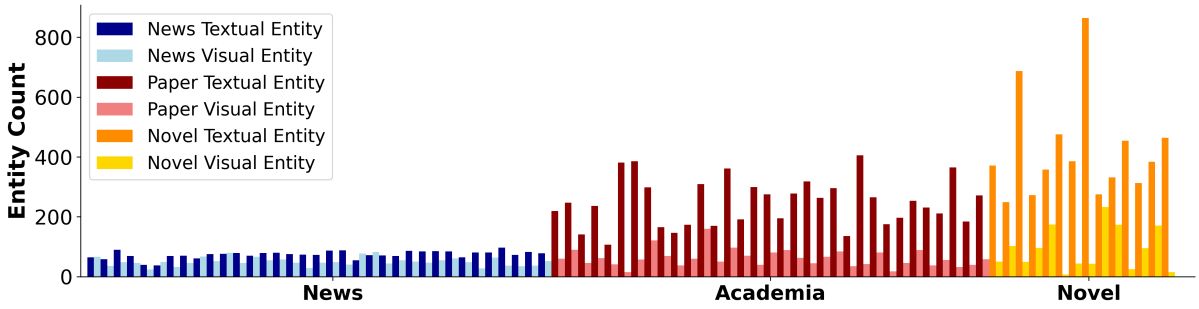

Figure 4: Entity Distribution Across Document Domains.

4.1.1 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 4. 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.

4.1.2 Experimental Setup and Results. We conduct a series of comparative experiments based on the CMEL dataset. The experiments cover three categories of unsupervised approaches: embedding-based methods (Emb), LLM-based methods (LLM), and our proposed SpecLink 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 [4], the LLM Qwen2.5-72B-Instruct [48], and the MLLM InternVL2.5-38B-MPO [56].

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. For each experiment, we report the average results over three runs. The outcomes proved to be highly consistent, with differences between runs not exceeding the order of 0.01. Overall, clustering-based methods significantly outperform the embedding-based and LLM-based methods in the CMEL task. Compared to other clustering-based methods, SpecLink achieves the best performance, improving micro-accuracy by about 15% and macro-accuracy by around 30%, clearly demonstrating its effectiveness in CMEL task.

4.2 Multimodal Document QA Experiments

We choose DocQA as the primary evaluation task for MMGraphRAG 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.

4.2.1 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 [37]: 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.

4.2.2 Comparison with Basic Methods. To ensure fair comparisons, eliminate potential biases arising from using the same model for both generation and evaluation, and assess the general applicability of our proposed prompt, this experiment selects various LLMs and MLLMs as comparison benchmarks. Although our experiments initially evaluated a broader set of three LLMs, namely Llama3.1-70B-Instruct [2], Qwen2.5-72B-Instruct [59], and Mistral-Large-Instruct-2411 [3], and three MLLMs, including Ovis1.6-Gemma2-27B [38], Qwen2-VL-72B [55], and InternVL2.5-38B-MPO [10], we focus our analysis on four representative models for clarity. Specifically, we present a comparison between Llama3.1(L) and Qwen2.5(Q) for LLMs, and between Qwen2(Qv) and InternVL2.5(Iv) for MLLMs.

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 [25] (NR): 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 (GR): The GraphRAG method was modified by removing the community detection part and using local mode querying to ensure fair comparison with other methods [14, 19]. 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 (MGR) over other methods.

The results for the DocBench and MMLongBench datasets are presented in Table 2 and Table 3, respectively. The following analysis from several key perspectives explains the superiority of MMGraphRAG.

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(MGR 60.7% vs. NR 43.6%; MGR 65.8% vs. NR 38.2% on DocBench). This indicates that MMGraphRAG 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.

Method	Type					Domain			Overall Acc.
	Aca.	Fin.	Gov.	Laws	News	Text.	Multi.	Una.	
LLM(L)	43.9	13.5	53.4	44.5	79.7	52.9	18.8	**81.5**	44.7
LLM(Q)	41.3	16.3	50.7	49.7	77.3	53.9	20.1	75.8	44.8
MLLM(Qv)	17.5	14.9	25.0	34.6	48.8	34.0	8.4	40.3	25.4
MLLM(Iv)	19.8	16.3	28.4	31.4	46.5	35.7	15.9	39.5	27.7
NR(L)	43.6	38.2	66.2	64.9	**80.2**	79.9	32.1	70.2	61.0
NR(Q)	43.6	34.4	62.8	65.4	75.0	81.6	30.5	67.7	59.5
GR(L)	40.6	27.1	56.8	59.7	75.0	73.5	24.4	76.6	54.7
GR(Q)	39.6	25.7	52.5	49.7	74.5	71.7	26.0	67.5	52.3
MGR(L-Qv)	51.8	59.4	62.8	60.7	77.9	79.1	77.8	70.2	74.0
MGR(Q-Qv)	51.8	62.9	**66.9**	68.6	76.2	**82.4**	81.1	67.7	75.2
MGR(L-Iv)	**60.7**	64.1	62.6	64.9	76.2	80.0	86.4	75.0	**77.5**
MGR(Q-Iv)	60.5	**65.8**	66.5	**70.4**	77.1	81.2	**88.7**	71.9	76.8

Table 2: Accuracy on DocBench

Method	Locations			Modalities					Overall	
	Sin.	Mul.	Una.	Cha.	Tab.	Txt.	Lay.	Fig.	Acc.	F1
LLM(L)	23.7	20.3	51.6	16.3	12.7	32.1	21.9	17.4	28.2	23.0
LLM(Q)	22.5	20.0	53.2	16.8	12.6	33.3	23.5	16.1	27.8	22.1
MLLM(Qv)	10.8	9.9	8.1	7.6	5.4	10.0	8.8	12.7	10.0	9.5
MLLM(Iv)	13.3	7.9	13.9	8.8	8.1	10.7	11.8	11.4	11.6	10.4
NR(L)	24.9	19.5	56.5	16.3	15.3	31.0	22.7	18.7	29.2	24.2
NR(Q)	22.3	16.4	52.5	15.1	10.8	30.3	20.3	14.9	26.2	20.9
GR(L)	16.3	12.3	**78.5**	7.6	6.7	25.1	15.0	10.6	27.2	18.2
GR(Q)	18.2	13.2	77.1	14.0	11.0	26.1	12.2	8.5	28.1	19.3
MGR(L-Qv)	37.6	13.8	55.2	26.4	29.2	26.4	13.8	21.2	32.6	28.1
MGR(Q-Qv)	38.7	20.1	51.6	26.2	30.0	29.3	**29.1**	29.2	34.8	30.4
MGR(L-Iv)	38.7	21.9	59.2	34.7	**36.6**	31.9	12.9	28.5	36.9	32.4
MGR(Q-Iv)	**39.6**	**26.7**	55.8	**34.7**	36.5	**33.8**	28.7	**34.6**	**38.8**	**34.1**

Table 3: Accuracy and F1 score on MMLongBench

Multimodal Information Processing Capability. Counterintuitively, results show that MLLMs alone do not outperform other methods on vision-centric queries, even worse than NaiveRAG based on image captioning. MMGraphRAG achieves a substantial improvement in multimodal query accuracy on DocBench (88.7%), far surpassing both the NaiveRAG (32.1%) and GraphRAG (26.0%), demonstrating its superior ability to retrieve and utilize visual knowledge. Similarly, on MMLongBench, MMGraphRAG attains the highest performance in charts, tables, figures and layout reasoning (34.7%, 36.6%, 29.1% and 34.6%), which are the most visually demanding tasks, confirming its effectiveness in fusing structural and semantic cues from complex multimodal inputs. These results validate that MMGraphRAG's explicit MMKG construction provide enhanced accuracy, setting a new standard for multimodal information processing.

4.3 Ablation Study

To investigate the individual contributions of our key components, we conduct a series of ablation studies. We aim to answer two primary questions: 1) How crucial is our sophisticated cross-modal fusion module? 2) What are the benefits of constructing an explicit MMKG compared to Multimodal RAG (MRAG) approaches?

4.3.1 Impact of the Cross-Modal Fusion Module. To validate the necessity of our fusion module, we design a simplified baseline, MMGraphRAG (w/o Fusion). In this variant, the cross-modal fusion module is replaced with a naive method that performs entity alignment based solely on embedding similarity. An image entity and a text entity are considered aligned if the cosine similarity of their embeddings exceeds a predefined threshold (0.7 in our experiments). We conduct this experiment on the academia (Aca) subset of DocBench, chosen for its high density of multimodal questions, making it an ideal testbed for evaluating fusion performance. The results are presented in Figure 5.

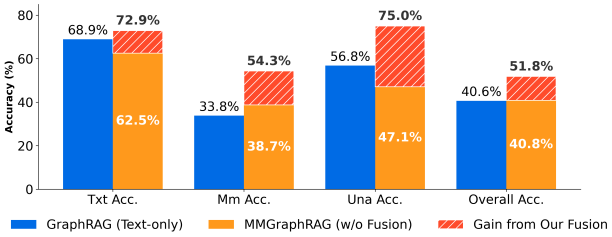

Figure 5: Impact of Cross-Modal Fusion Module by Metric

The analysis of the results reveals several key insights:

- **Textual Performance:** Interestingly, the naive fusion approach degrades performance on text-only (Txt) questions. This suggests that imprecise entity linking introduces noise and irrelevant visual context, which can disrupt purely textual reasoning. In contrast, our full model enhances text accuracy, indicating that well-aligned visual information can effectively supplement textual context, clarifying ambiguities and providing missing details.
- **Multimodal Performance:** On multimodal (Mm) questions, simply incorporating the image KG with a naive alignment yields a marginal improvement (+4.9%) over the text-only GraphRAG. However, our full MMGraphRAG framework achieves a substantial gain (+21.5%). This demonstrates that the fusion method is essential for building precise and meaningful cross-modal links, which are critical for answering complex multimodal queries.
- **Hallucination Suppression:** The most significant difference is observed in unanswerable (Una) questions. The naive fusion method's accuracy drops sharply (-9.7%), showing that introducing poorly aligned visual information acts as noise and exacerbates model hallucinations. Conversely, our method's robust fusion process strengthens the logical coherence between modalities, improving the model's ability to identify unanswerable questions by a remarkable 18.2%. This confirms that our fusion method builds authentic and valuable connections rather than spurious correlations.

These results collectively underscore that our proposed cross-modal fusion module is a critical component, not just an incremental improvement.

4.3.2 Comparison with MRAG Methods. To isolate the benefits of our structured MMKG, we compare MMGraphRAG against a representative MRAG framework, M3DOCRAG (M3DR). This comparison effectively serves as an ablation of our explicit graph-building pipeline, contrasting it with approaches that typically project multimodal features into a shared embedding space for retrieval. The experiment was conducted on the MMLongBench dataset with Qwen2 7b series, and the results are shown in Table 4.

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
MGR	34.3	12.5	**35.1**	48.2	24.6	18.2	22.2	**26.5**	**23.8**

Table 4: M3DOCRAG vs MMGraphRAG performance

The results highlight the advantages of our MMKG-based approach. By explicitly modeling relationships between visual and textual entities, MMGraphRAG more effectively handles queries requiring deep semantic alignment between images and text. This is particularly evident in its superior performance on questions related to charts and tables (C.T.), where it outperforms M3DR by a significant margin (48.2% vs. 39.0%).

Furthermore, ablating our explicit graph structure in favor of an MRAG approach reveals a critical weakness in handling unanswerable (Una) questions. MRAG methods, which rely on a unified embedding space, can be misled by noisy but semantically adjacent evidence, causing the model to confidently generate incorrect answers. In contrast, the structured reasoning enabled by our MMKG allows the model to more reliably traverse evidence and assess its completeness. This structured approach significantly reduces the generation of misleading answers and enhances the model's robustness, as shown by the massive performance gap on unanswerable questions (MMGR 35.1% vs. M3DR 5.8%).

This analysis validates that the explicit construction of an MMKG is a crucial design choice for achieving robust and accurate multimodal DocQA, particularly in complex scenarios requiring nuanced, cross-modal reasoning.

5 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

[1] Shayan A. Tabrizi, Azadeh Shakery, Masoud Asadpour, Maziar Abbasi, and Mohammad Ali Tavallaie. 2013. Personalized PageRank Clustering: A graph clustering algorithm based on random walks. *Physica A: Statistical Mechanics and its Applications* 392, 22 (2013), 5772–5785. doi:10.1016/j.physa.2013.07.021

[2] Meta AI. 2024. LLaMA-3.1-70B-Instruct. Available at https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

[3] Mistral AI. 2024. Mistral-Large-Instruct-2411. Available at https://huggingface.co/mistralai/Mistral-Large-Instruct-2411.

[4] Stability AI. 2024. stella-en-1.5B-v5. https://huggingface.co/stabilityai/stella-en-1.5B-v5. Open-weight English language model.

[5] Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez de Lacalle. 2025. Vision-Language Models Struggle to Align Entities across Modalities. arXiv:2503.03854 [cs.CL] https://arxiv.org/abs/2503.03854

[6] Chenyang Bu, Guojie Chang, Zihao Chen, Cunyuan Dang, Zhize Wu, Yi He, and Xindong Wu. 2025. Query-Driven Multimodal GraphRAG: Dynamic Local Knowledge Graph Construction for Online Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*. 21360–21380.

[7] Guikun Chen, Jin Li, and Wenguan Wang. 2024. Scene Graph Generation with Role-Playing Large Language Models. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 132238–132266. https://proceedings.neurips.cc/paper_files/paper/2024/file/ee74a6ade401e200985e2421b20bbae4-Paper-Conference.pdf

[8] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).

[9] Yixiang Chen, Penglei Sun, Xiang Li, and Xiaowen Chu. 2025. MRD-RAG: Enhancing Medical Diagnosis with Multi-Round Retrieval-Augmented Generation. arXiv:2504.07724 [cs.CL] https://arxiv.org/abs/2504.07724

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. 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.

[11] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. *CoRR* abs/2411.04952 (2024). https://doi.org/10.48550/arXiv.2411.04952

[12] Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. 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*. https://openreview.net/forum?id=A4RH0kNxpn

[13] Dingsheng Deng. 2020. DBSCAN Clustering Algorithm Based on Density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*. 949–953. doi:10.1109/IFEEA51475.2020.00199

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

[15] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130 [cs.CL] https://arxiv.org/abs/2404.16130

[16] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=ogjBpZ8uSi

[17] Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal Entity Linking: A New Dataset and A Baseline. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 993–1001. doi:10.1145/3474085.3475400

[18] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 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*. Springer, 986–996.

[19] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779* (2024).

[20] Mahd Hindi, Linda Mohammed, Ommama Maaz, and Abdulmalik Alwarafy. 2025. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access* 13 (2025), 46171–46189. doi:10.1109/ACCESS.2025.3550145

[21] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[22] Hongjie Jia, Shifei Ding, Xinzheng Xu, and Ru Nie. 2014. The latest research progress on spectral clustering. *Neural Computing and Applications* 24 (2014), 1477–1486.

[23] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and S. Sarasvady. 2014. DBSCAN: Past, present and future. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*. 232–238. doi:10.1109/ICADIWT.2014.6814687

[24] Trupti M Kodinariya, Prashant R Makwana, et al. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal* 1, 6 (2013), 90–95.

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[27] Wanying Liang, Pasquale De Meo, Yong Tang, and Jia Zhu. 2024. A Survey of Multi-modal Knowledge Graphs: Technologies and Trends. *ACM Comput. Surv.* 56, 11, Article 273 (June 2024), 41 pages. doi:10.1145/3656579

[28] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. 2003. The global k-means clustering algorithm. *Pattern Recognition* 36, 2 (2003), 451–461. doi:10.1016/S0031-3203(02)00060-2 Biometrics.

[29] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 22820–22840. https://proceedings.neurips.cc/paper_files/paper/2023/file/47393e8594c82ce8fd83adc672cf9872-Paper-Conference.pdf

[30] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. 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)*.

[31] Zihan Ling, Zhiyao Guo, Yixuan Huang, Yi An, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. MMKB-RAG: A Multi-Modal Knowledge-Based Retrieval-Augmented Generation Framework. arXiv:2504.10074 [cs.AI] https://arxiv.org/abs/2504.10074

[32] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A Survey on Hallucination in Large Vision-Language Models. *CoRR* abs/2402.00253 (2024). https://doi.org/10.48550/arXiv.2402.00253

[33] Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Botian Shi, and Ding Wang. 2025. Aligning Vision to Language: Text-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning. arXiv:2503.12972 [cs.CV] https://arxiv.org/abs/2503.12972

[34] Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. arXiv:2504.12330 [cs.CL] https://arxiv.org/abs/2504.12330

[35] Qi Liu, Yongyi He, Tong Xu, Defu Lian, Che Liu, Zhi Zheng, and Enhong Chen. 2024. 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* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 1909–1919. doi:10.1145/3627673.3679793

[36] Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024. OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting. *arXiv preprint arXiv:2410.07549* (2024).

[37] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* (2023).

[38] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797* (2024).

[39] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2025. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems* 37 (2025), 95963–96010.

[40] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A Survey of Multimodal Retrieval-Augmented Generation. arXiv:2504.08748 [cs.IR] https://arxiv.org/abs/

2504.08748
[41] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *CoRR* abs/2004.13637 (2020). https://arxiv.org/abs/2004.13637
[42] Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. 2023. Entity Linking Meets Deep Learning: Techniques and Solutions. *IEEE Transactions on Knowledge and Data Engineering* 35, 3 (2023), 2556–2578. doi:10.1109/TKDE.2021.3117715
[43] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2015), 443–460. doi:10.1109/TKDE.2014.2327028
[44] Kristina P. Sinaga and Miin-Shen Yang. 2020. Unsupervised K-Means Clustering Algorithm. *IEEE Access* 8 (2020), 80716–80727. doi:10.1109/ACCESS.2020.2988796
[45] Shezheng Song, Shan Zhao, ChengYu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. 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 (Mar. 2024), 19008–19016. doi:10.1609/aaai.v38i17.29867
[46] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. 2021. Energy-Based Learning for Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13936–13945.
[47] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation From Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[48] Qwen Team. 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.
[49] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* 9, 1 (2019), 1–12.
[50] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 6558–6569. doi:10.18653/v1/P19-1656
[51] Ultralytics. 2023. Ultralytics YOLOv8: Cutting-Edge Object Detection Models. https://github.com/ultralytics/ultralytics. Accessed: 2025-07-14.
[52] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3156–3164. doi:10.1109/CVPR.2015.7298935
[53] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839* (2024).
[54] Jingyi Wang, Jianzhong Ju, Jian Luan, and Zhidong Deng. 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. doi:10.1109/ICASSP49660.2025.10887586
[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
[56] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, Jifeng Dai, and et al. 2024. InternVL2.5-MPO: Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. arXiv preprint arXiv:2411.10442.
[57] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=s5epFPdIW6
[58] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[59] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).
[60] Barry Menglong Yao, Yu Chen, Qifan Wang, Sijia Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. 2023. AMELI: Enhancing Multimodal Entity Linking with Fine-Grained Attributes. *CoRR* abs/2305.14725 (2023). https://doi.org/10.48550/arXiv.2305.14725
[61] Chen Yin and Zixuan Zhang. 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)*. Atlantis Press, 677–684.
[62] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. *CoRR* abs/2410.10594 (2024). https://doi.org/10.48550/arXiv.2410.10594
[63] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene Graph Parsing With Global Context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[64] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and Stella: distillation of SOTA embedding models. *arXiv preprint arXiv:2412.19048* (2024).
[65] Fengzhi Zhao, Zhezhou Yu, Tao Wang, and Yi Lv. 2024. Image Captioning Based on Semantic Scenes. *Entropy* 26, 10 (2024). doi:10.3390/e26100876
[66] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. 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.
[67] Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics* 8 (2020), 109–124.
[68] Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *arXiv preprint arXiv:2407.10701* (2024).

A Availability of Data and Materials

A complete account of all parameters and prompts employed in this study has been omitted from the main text and appendix owing to manuscript length limitations. To ensure full reproducibility, the entire dataset, associated source code, and step-by-step replication instructions will be released publicly upon acceptance of this paper. In the interim, these resources are provided for peer review purposes at the following anonymous repository: https://anonymous.4open.science/r/CMEL-dataset-CF7D.

B A Detailed Introduction to the CMEL Dataset

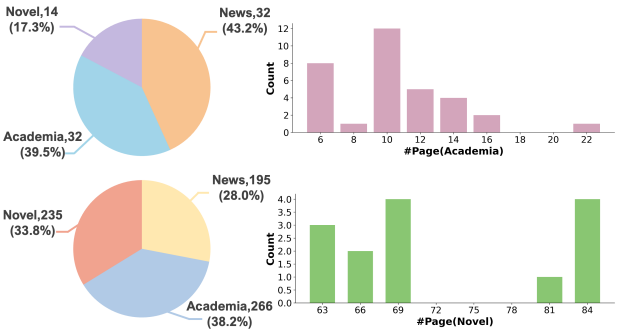

Figure 6: The Distribution of CMEL dataset. In the top-left, the number and proportion of documents in each domain are shown; in the bottom-left, the number and proportion of images in each domain are displayed; in the top-right, the page distribution of academia domain documents is provided; and in the bottom-right, the page distribution of novel documents is shown. All news domain documents are one page.

The 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

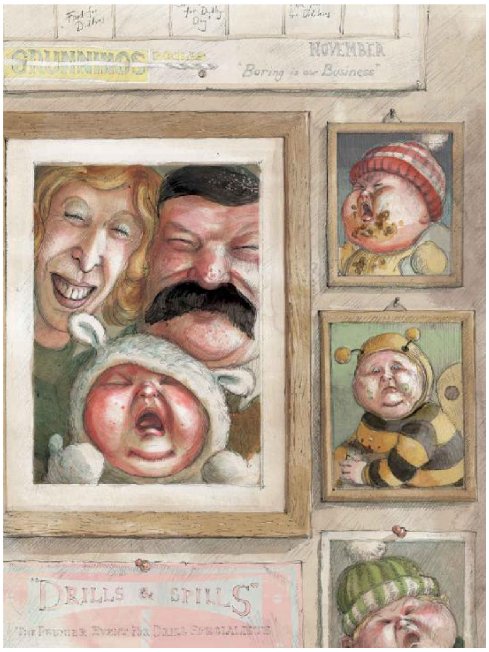

Figure 7: The Dursleys' photo wall. From Chapter 1 of "Harry Potter and the Sorcerer's Stone".

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. 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 C.

The distribution of dataset documents and images is shown in Figure 6. 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[53] 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 7 as a specific example.

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

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

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.

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

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.

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"
            ]
        },
        ...
    ]
```

C 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[68]. 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 JSON file named "kv store chunk knowledge".
- **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 JSON file named "merged entities", which serves as guides for the next step.

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 description 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.

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.

D 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[61] (MLM), bge-m3[8] (BGE), and stella-en-1.5B-v5[64] (Ste). For the LLMs, we chose Llama3.1-70B-Instruct[2] (L) and Qwen2.5-72B-Instruct[59] (Q), while for the MLLMs, we selected Qwen2-VL-72B[55] (Qv) and InternVL2.5-38B-MPO[10] (Iv).

After clustering, CMEL requires selecting the appropriate cluster for the target image entity, with two specific methods: KNN[18] and LLM-based judgment.

Meth.	**micro/macro Acc.**			**Overall.**
	News	Aca.	Nov.	
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
Ste	10.8/8.4	33.1/34.5	9.0/7.5	20.0/16.8
L-Qv	10.8/8.4	33.1/34.5	9.0/7.5	20.0/16.8
L-Iv	10.8/16.7	30.2/30.0	13.5/13.3	20.8/16.8
Q-Qv	31.2/24.1	32.2/33.3	19.4/23.2	26.1/26.8
Q-Iv	33.3/24.1	36.8/36.1	17.4/20.8	27.1/27.0
DB	48.4/43.1	57.0/58.4	29.9/31.3	43.5/44.3
KM	48.4/41.5	58.2/59.4	29.6/29.4	43.9/43.4
PR	50.5/43.0	61.0/56.4	29.2/33.2	44.7/44.2
Lei	50.5/42.1	66.7/64.3	30.4/37.2	47.7/47.9
Spec	**57.5/50.9**	**70.1/66.1**	**31.0/39.8**	**49.7/55.1**

Table 6: Complete Results for CMEL dataset

The main text reports results based on the LLM-based judgment, while Table 6 in the Appendix presents the complete results obtained using the KNN method for reference.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009