Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, No. 6, December, 2025

Enhance Multimodal Retrieval-Augmented Generation Using Multimodal Knowledge Graph

Shue-Kei How [1], Lee-Yeng Ong [1*], Meng-Chew Leow [1]

[1] *Faculty of Information Science and Technology (FIST), Multimedia University, Melaka 75450, Malaysia.*

Abstract

Large Language Models (LLMs) have shown impressive capabilities in natural language understanding and generation tasks. However, their reliance on text-only input limits their ability to handle tasks that require multimodal reasoning. To overcome this, Multimodal Large Language Models (MLLMs) have been introduced, enabling inputs such as images, text, video and audio. While MLLMs address some limitations, they often suffer from hallucinations because of over-reliance on internal knowledge and face high computational costs. Traditional vector-based multimodal RAG systems attempt to mitigate these issues by retrieving supporting information, but often suffer from cross-modal misalignment, where independently retrieved text and image content cannot align meaningfully. Motivated by the structured retrieval capabilities of text-based knowledge graph RAG, this paper proposes VisGraphRAG to address the challenge by modelling structured relationships between images and text within a unified MMKG. This structure enables more accurate retrieval and better alignment across modalities, resulting in more relevant and complete responses. The experimental results show that VisGraphRAG significantly outperforms the vector database-based baseline RAG, achieving a higher answer accuracy of 0.7629 compared to 0.6743. Besides accuracy, VisGraphRAG also shows superior performance in key RAGAS metrics such as multimodal relevance (0.8802 vs 0.7912), showing its stronger ability to retrieve relevance information across modalities. These results underscore the effectiveness of the proposed Multimodal Knowledge Graph (MMKG) methods in enhancing cross-modal alignment and supporting more accurate, context-aware generation in complex multimodal tasks.

Keywords:

Multimodal Retrieval-Augmented Generation;
Multimodal Knowledge Graph;
Multimodal Large Language Models;
Cross-Modality Alignment.

Article History:

Received:	07	June	2025
Revised:	19	November	2025
Accepted:	24	November	2025
Published:	01	December	2025

1- Introduction

Large Language Models (LLMs) are powerful language processing systems with billions of parameters, capable of performing a wide range of Natural Language Processing (NLP) tasks such as question answering, problem-solving, information retrieval and so on [1]. However, traditional LLMs are limited to processing plain text inputs and are not capable of tasks that require multimodal understanding [2].

To overcome this limitation, Multimodal Large Language Models (MLLMs) such as those proposed in [3–5] have been developed to handle inputs across various modalities such as text, images, video, and audio. While MLLMs have shown promising results, they still face two key challenges, including a tendency to produce hallucinations because of over-reliance on internal knowledge, and high computational cost in training and fine-tuning, caused by their massive parameter sizes [6].

Consequently, Multimodal Retrieval-Augmented Generation (RAG) has emerged as a scalable alternative, allowing MLLMs to query external knowledge sources and ground their responses in retrieved evidence. Existing multimodal

[*] **CONTACT**: lyong@mmu.edu.my

DOI: http://dx.doi.org/10.28991/ESJ-2025-09-06-025

© 2025 by the authors. Licensee ESJ, Italy. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC-BY) license (https://creativecommons.org/licenses/by/4.0/).

RAG frameworks, including [7–9] typically follow a two-step process: first, retrieving semantically similar information from a vector database; and second, generating responses guided by the retrieved content. However, these vector-based systems often treat each modality independently and rely solely on similarity-based retrieval, without explicitly capturing or leveraging the semantic relationships between modalities [10]. As a result, they frequently suffer from cross-modal misalignment, where information from different modalities (e.g., text and image) fails to align or correspond meaningfully during retrieval or reasoning [11].

To improve alignment and interpretability, some studies have explored Knowledge Graph-based RAG frameworks [12, 13], but these are restricted to text-only scenarios. A knowledge graph (KG) is a structured representation where nodes represent entities or concepts and edges capture their relationships [14]. Unlike vector databases, knowledge graphs organize information as interconnected triples of entities and relations, offering a more interpretable and context-aware retrieval process. While KG-RAG has improved reasoning and factual consistency in text-only settings, it remains restricted to unimodal contexts and lacks the ability to model cross-modal relationships essential for multimodal understanding.

Inspired by the structure and benefits of text-only KG-RAG, this study proposes VisGraphRAG, a novel multimodal RAG framework that integrates a Multimodal Knowledge Graph (MMKG) as its external knowledge source to address the issue of cross-modal misalignment. MMKGs extend the KG concept by incorporating diverse modalities such as linking text-based entities to images [15]. Previous studies such as [10, 16, 17], have primarily focused on leveraging MMKGs to understand cross-modal relationships and improve retrieval accuracy. Although these works are not directly applied to multimodal RAG, they demonstrate the potential of MMKGs to enhance cross-modal alignment. Despite this promise, there remains a lack of research investigating how MMKG integration affects the overall performance of multimodal RAG systems, particularly in terms of retrieval quality and response generation. Therefore, the key contributions of this study are as follows:

- Introducing VisGraphRAG, a framework that improves cross-modal alignment in multimodal RAG by using a structured MMKG.
- Evaluating how the integration of MMKG affects the performance of multimodal RAG using RAGAS metrics.

The organization of this study is as follows. Section 2 reviews related topics on multimodal RAG, identifies key limitations in existing approaches, and explores the potential of MMKGs as a solution. Section 3 details the methodology used to develop the proposed VisGraphRAG framework. Section 4 presents the experimental setup, detailing how the experiments were designed and conducted to evaluate the effectiveness of the proposed method. Section 5 discusses the results and key observations. Finally, Section 6 concludes the study and outlines potential future directions.

2- Related Works

2-1- MLLM

Multimodal Large Language Models (MLLMs) are typically developed by extending Large Language Models (LLMs), with additional capabilities to process and understand various types of data. By working with different types of data, MLLMs strive to develop a deeper and more refined understanding of the world, like how humans perceive and interpret diverse forms of information. MLLM typically are composed of several key components, including a modality encoder that extracts features from different input types, an input projector that aligns non-text features with the text embedding space, and an LLM backbone that performs reasoning over the unified representation. The output projector converts the model's internal representations into modality-specific signals, which are then used by the modality generator to produce the final output. This architecture allows MLLMs to process a variety of input modalities and generate outputs that are aligned with the target modality [18].

Despite their strong performance, MLLMs rely heavily on knowledge encoded in their parameters, which can lead to issues such as outdated information, hallucinations, and limited interpretability [18]. Hallucinations refer to content that is inaccurate or illogical relative to the input source [19]. To address these limitations, multimodal RAG has emerged as a promising solution.

2-2- Multimodal RAG

Unlike static models such as MLLMs, which rely solely on pre-trained data, RAG systems can actively search and incorporate relevant information from external sources, making them more adaptable and capable of addressing questions that require the latest and context-specific knowledge. Recent studies, including [7-9], have explored the development of multimodal RAG systems.

Although these studies differ in how they handle multimodal alignment, they share a common retrieval strategy. For example, Chen et al. [7] employs a fused multimodal encoder (ViT + T5) to jointly encode image-text pairs into a shared embedding space, which is then stored in a vector database. Riedler & Langer [8] explores two approaches: the first encodes images and text separately and stores them in distinct vector databases; the second converts images into textual summaries and stores both the image summaries and text chunks in a shared vector store. In contrast, Joshi et al. [9] generates a summary for each page image using an MLLM and stores the resulting image-summary embeddings in one vector database, while textual content is stored separately in another.

Despite differences in fusion strategies and preprocessing methods, all three approaches ultimately follow the same underlying architecture: they encode multimodal data (images and text) into dense embeddings and store them in vector databases for similarity-based retrieval. When a user submits a query, it is also converted into an embedding and used to perform a semantic search in the vector database. During the retrieval phase, the system selects the top-k most relevant entries based on similarity scores, with k being a predefined parameter. The retrieved information is then passed to the generator (typically a MLLM) which synthesizes a coherent and contextually appropriate response.

While efficient, this vector-based approach has important limitations. In such systems, the retrieval is performed based on semantic similarity between the query and the stored data. However, this process does not capture explicit relationships between entities across different modalities [12]. Moreover, text and images inherently encode and structure information in different ways, which leads to a modality gap [10]. Without a structured understanding of how these modalities relate to one another, the model finds it difficult to reason over complex and interconnected data. Furthermore, cross-modal misalignment may occur, where the retrieved text may not meaningfully relate to the image. This limitation can negatively impact the accuracy and completeness of both the retrieval process and the final generated response [16]. Therefore, there is a need for approaches that go beyond semantic similarity by introducing structured relationships between multimodal data.

2-3- Knowledge Graph-Based RAG

To go beyond semantic similarity, several methods such as [12, 13] have proposed integrating knowledge graphs into RAG systems to enhance reasoning and retrieval performance. Instead of relying solely on vector databases, these approaches leverage knowledge graphs to introduce structured relationships between textual entities. A knowledge graph (KG) is a structured, directed, and labelled graph where nodes indicate entities or concepts, and edges illustrate the relationships between them [20]. Beyond this graphical representation, a KG can also be viewed as an organized collection of facts expressed as factual triples in the form of (head, predicate, tail) under the RDF framework. Here, the "head" and "tail" signify entities, while the "predicate" denotes the type of relationship between them [21].

By integrating a knowledge graph into the RAG framework, these systems improve the interpretability and factual grounding of generated outputs. The retrieval stage can access graph-structured data rather than flat and unconnected text embeddings. This allows the model to better understand how concepts relate and to retrieve more contextually accurate and logically consistent evidence. However, most existing knowledge graph-based RAG methods are designed for text-only scenarios. They do not account for relationships that span across different modalities such as images and text. Nevertheless, the structural advantages of knowledge graphs have inspired recent efforts to explore the integration of Multimodal Knowledge Graphs (MMKGs) into the multimodal RAG pipeline.

2-4- MMKG

When a knowledge graph integrates diverse modalities such as text, images, audio, or video, it becomes a multimodal knowledge graph (MMKG). These different modalities provide complementary perspectives on the same entity, enriching the graph's representation and enhancing both knowledge graph applications and the ability of machines to interpret complex real-world data [21].

Recent studies such as [10, 16, 17] have leveraged MMKG to better understand the relationships of cross-modal data and improve retrieval accuracy. Although these studies are not directly focused on the multimodal RAG domain, they demonstrate that MMKG is an effective approach for enhancing cross-modality understanding and alignment. MMKGs have the potential to bridge the modality gap and enhance multimodal RAG systems by explicitly modeling cross-modal relationships. This opens new possibilities for more accurate, explainable, and semantically aligned multimodal generation.

3- Proposed Method

This study proposes an MMKG-based RAG system called VisGraphRAG. While it follows the general architecture of a typical multimodal RAG framework, VisGraphRAG introduces a key difference in the retrieval phase. Instead of retrieving image and text information separately from a vector database, VisGraphRAG retrieves data from a MMKG. From a theoretical perspective, the use of MMKG enables more interpretable and semantically coherent reasoning by explicitly modelling entities and their relationships as graph structures. Unlike vector-based retrieval, which retrieves modalities independently based on similarity scores, MMKG-based retrieval additionally leverages relational paths and explicitly defined connections between modalities. This ensures that related images and text remain tightly connected, thereby reducing the risk of cross-modal misalignment. As a result, the system is better able to retrieve accurate and contextually aligned multimodal information and ultimately leading to more accurate and grounded responses. Figure 1 illustrates the difference between the traditional vector-based multimodal RAG and the proposed method. The figure highlights how MMKG-based retrieval preserves consistency between visual and textual information, reducing the likelihood of mismatched content and helping prevent confusion during response generation. A detailed analysis of this improvement is provided in the discussion section.

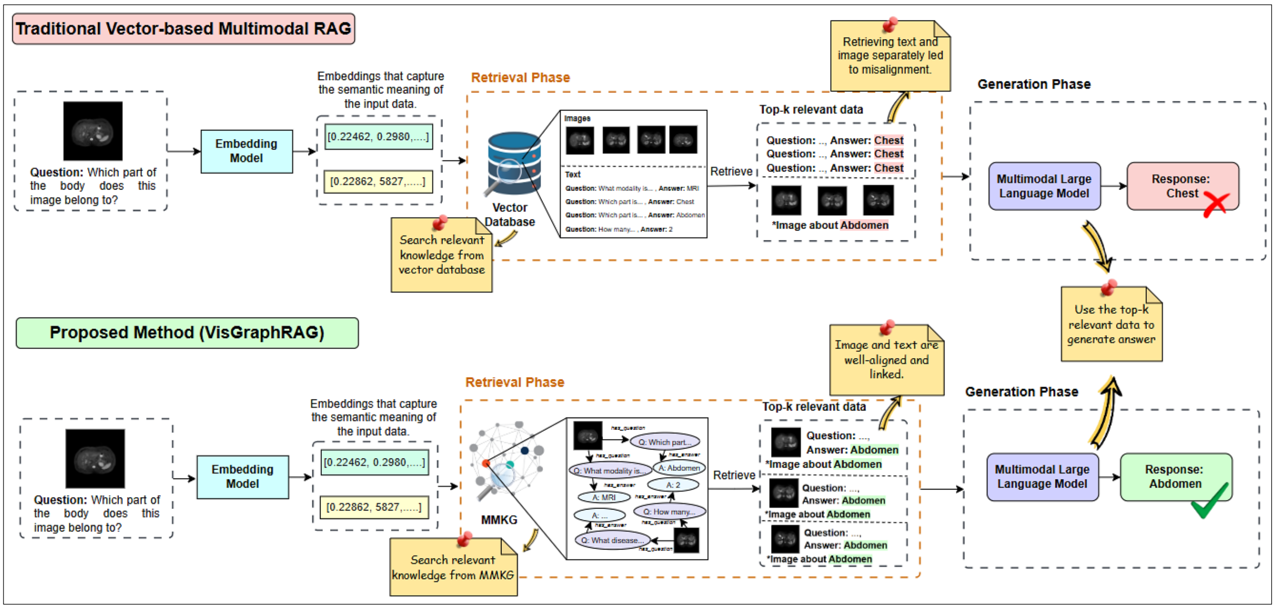

Traditional Vector-based Multimodal RAG
Question: Which part of the body does this image belong to?
Embedding Model
Embeddings that capture the semantic meaning of the input data.
[0.22462, 0.2980,....]
[0.22862, 5827,.....]
Retrieval Phase
Vector Database
Search relevant knowledge from vector database
Images
Text
Question: What modality is..., Answer: MRI
Question: Which part is..., Answer: Chest
Question: Which part is..., Answer: Abdomen
Question: How many..., Answer: 2
Retrieve
Retrieving text and image separately led to misalignment.
Top-k relevant data
Question: ..., Answer: Chest
Question: ..., Answer: Chest
Question: ..., Answer: Chest
*Image about Abdomen
Generation Phase
Multimodal Large Language Model
Response: Chest
Use the top-k relevant data to generate answer
Proposed Method (VisGraphRAG)
Question: Which part of the body does this image belong to?
Embedding Model
Embeddings that capture the semantic meaning of the input data.
[0.22462, 0.2980,....]
[0.22862, 5827,.....]
Retrieval Phase
MMKG
Search relevant knowledge from MMKG
has_question
Q: Which part...
has_answer
A: Abdomen
Q: What modality is...
A: MRI
A: 2
A: ...
Q: How many...
Q: What disease...
Retrieve
Image and text are well-aligned and linked.
Top-k relevant data
Question: ..., Answer: Abdomen
*Image about Abdomen
Question: ..., Answer: Abdomen
*Image about Abdomen
Question: ..., Answer: Abdomen
*Image about Abdomen
Generation Phase
Multimodal Large Language Model
Response: Abdomen

Figure 1. Comparison of Traditional Vector-based Multimodal RAG and VisGraphRAG

Building on this comparison, the following sections present the development of the proposed VisGraphRAG framework. As illustrated in Figure 2, the overall architecture consists of two main phases: MMKG construction and RAG integration. The methodology for each phase is detailed in the subsequent sections.

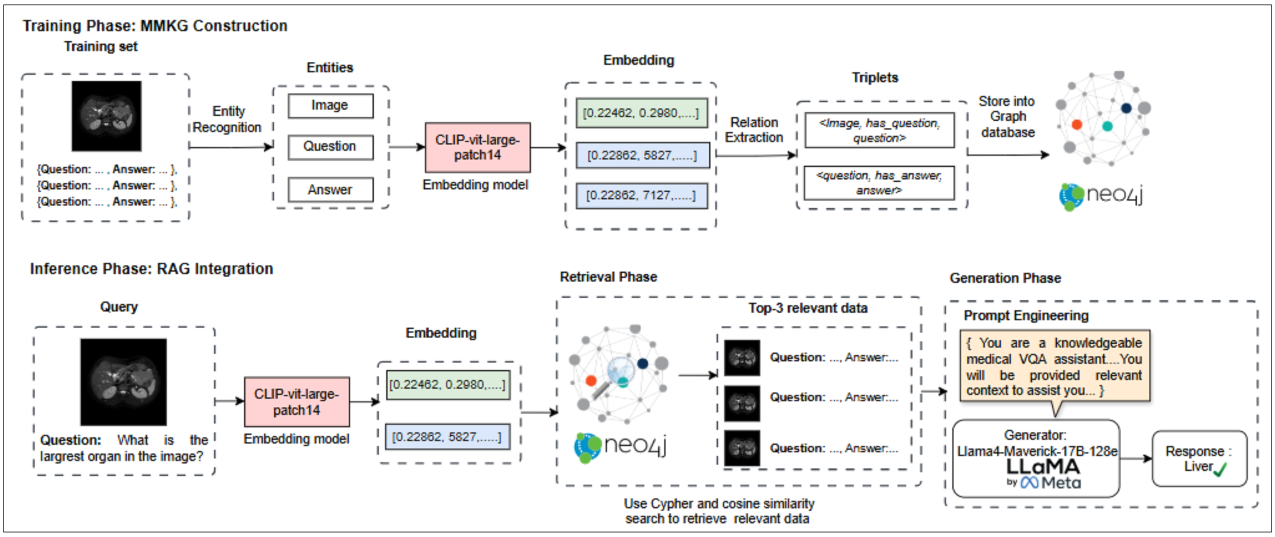
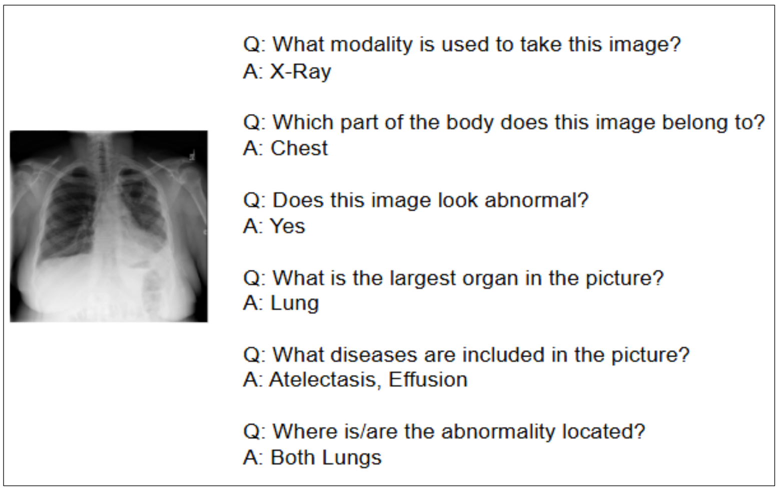

Training Phase: MMKG Construction
Training set
{Question: ..., Answer: ...},
{Question: ..., Answer: ...},
{Question: ..., Answer: ...},
Entity Recognition
Entities
Image
Question
Answer
CLIP-vit-large-patch14
Embedding model
Embedding
[0.22462, 0.2980,....]
[0.22862, 5827,.....]
[0.22862, 7127,.....]
Relation Extraction
Triplets
<Image, has_question, question>
<question, has_answer, answer>
Store into Graph database
neo4j
Inference Phase: RAG Integration
Query
Question: What is the largrest organ in the image?
CLIP-vit-large-patch14
Embedding model
Embedding
[0.22462, 0.2980,....]
[0.22862, 5827,.....]
Retrieval Phase
neo4j
Top-3 relevant data
Question: ..., Answer:...
Question: ..., Answer:...
Question: ..., Answer:...
Use Cypher and cosine similarity search to retrieve relevant data
Generation Phase
Prompt Engineering
{ You are a knowledgeable medical VQA assistant....You will be provided relevant context to assist you... }
Generator: Llama4-Maverick-17B-128e
LLaMA by Meta
Response : Liver

Figure 2. Design of VisGraphRAG

3-1- Dataset

This study utilizes the SLAKE (Semantically Labelled Knowledge-Enhanced) dataset, introduced in Liu et al. [22], which is specifically designed for Medical Visual Question Answering (Med-VQA). SLAKE is a bilingual dataset (English and Chinese) comprising 642 expert-annotated medical images, primarily sourced from X-rays, CT scans, and MRIs, along with a total of 14,000 question-answer (QA) pairs. The QA pairs cover a broad spectrum of medical content, including organ localization, image modality, visual abnormalities, and pathological indications. The dataset spans various anatomical regions such as the head, neck, chest, and abdomen. To provide readers with a clearer view of the dataset structure, Figure 3 presents a sample medical image along with several of its associated QA pairs.



Q: What modality is used to take this image?
A: X-Ray

Q: Which part of the body does this image belong to?
A: Chest

Q: Does this image look abnormal?
A: Yes

Q: What is the largest organ in the picture?
A: Lung

Q: What diseases are included in the picture?
A: Atelectasis, Effusion

Q: Where is/are the abnormality located?
A: Both Lungs

Figure 3. Example SLAKE image with associated medical QA pairs

3-2- MMKG Construction

The construction of the Multimodal Knowledge Graph (MMKG) begins with entity recognition and relation extraction, which aims to identify the key components (entities) within the dataset and define the semantic relationships that connect them. This process transforms unstructured multimodal data into a structured graph format, where nodes represent distinct pieces of information and edges capture their interconnections.

In this study, each image, question, and answer are treated as a distinct entity and modelled as an individual node in the graph. For each image, a has_question relationship is established, linking the image node to the questions associated with it. In other words, the image serves as the central node, and each question related to the image is connected as a separate node. In triple format, this is represented as <image, has_question, question>. In turn, each question node is linked to its corresponding answer node via the has_answer relationship, showing the connection between questions and their respective answers. As a result, the relationships are represented in the form <question, has_answer, answer>. This structure facilitates more interpretable and knowledge-rich representations for reasoning tasks. To support semantic similarity computation, each image and textual element (questions and answers) is encoded into a numerical embedding vector that captures its contextual meaning. These embeddings are stored as properties within their respective nodes, enabling efficient vector-based retrieval and similarity scoring during inference.

Once the entities, relationships, and embeddings are defined, the entire MMKG is stored in Neo4j, a high-performance graph database designed for storing and querying large-scale graph structures. Figure 4 illustrates how the multimodal data is structured within Neo4j.

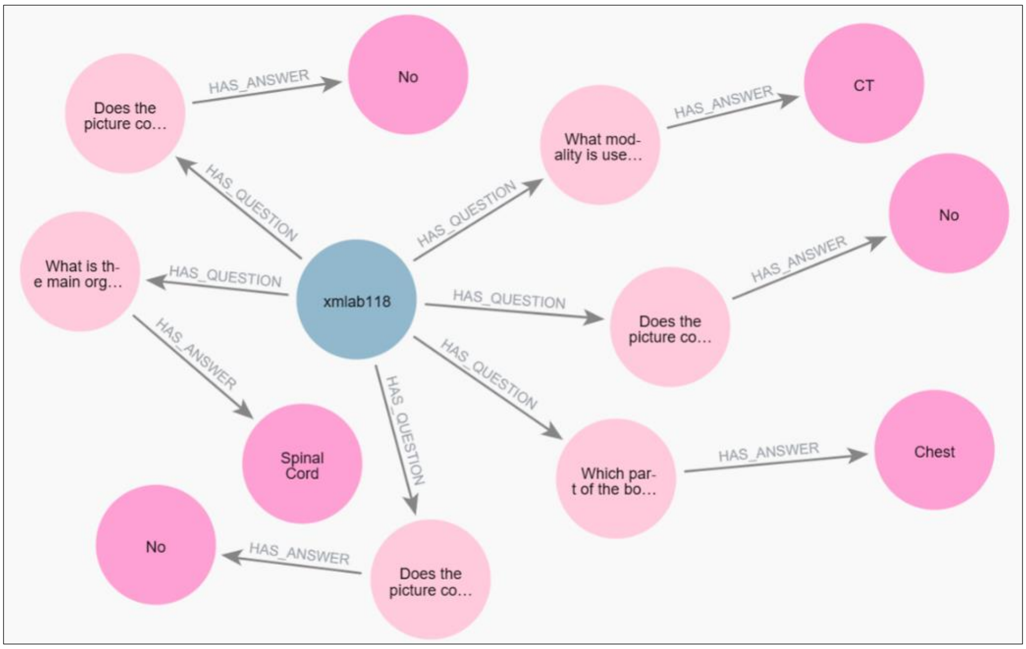


Figure 4. Example illustrating how data is stored in Neo4j

3-3- RAG Integration

After constructing the MMKG, it is now ready to serve as an external database, enabling its integration with the multimodal RAG system.

In the retrieval phase, the primary aim is to efficiently access relevant data from the MMKG to support the generative model in producing accurate and contextually relevant responses. Since the data is stored in the Neo4j graph database, the retrieval process leverages a hybrid mechanism that combines Cypher queries for structural graph traversal with cosine similarity computations for semantic alignment. Initially, Cypher was used to retrieve candidate nodes that are structurally connected. Then, embedding vectors associated with these candidates are compared using cosine similarity to identify the most semantically relevant entities. This integration enables the system to return contextually appropriate nodes even when the textual descriptions differ in wording or when the input image is visually similar but not identical to the stored data, thereby enhancing retrieval flexibility.

Cosine similarity search works by leveraging embeddings stored in the graph database. The equation for cosine similarity is defined as Equation 1 [23], where A represents the weights of the features in vector A, B represents the weights of the characteristics in vector B, i denotes the index of each feature dimension, and n represents the total number of dimensions in the embedding vectors. The cosine similarity measures the cosine of the angle between two vectors. A smaller cosine angle indicates higher similarity, as the vectors are more aligned.

$$Cos\ \alpha = \frac{A \times B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{1}$$

Following the retrieval phase, the system proceeds to the generation phase, where a MLLM is employed as the generator within the RAG framework. In this study, prompt engineering is utilized to ensure that the generated answers align closely with the task-specific requirements. The prompt provided to the MLLM is carefully crafted to guide its behavior:

> *"You are a highly knowledgeable medical visual question answering assistant. Your task is to analyse medical images and answer the question. You will be provided with relevant context to assist you. Keep your response concise and straight to the point. Do not add any information or explanation that is not supported by the context. Phrase even short answers as full sentences that restate the question for clarity."*

This structured prompt ensures the MLLM remains focused, avoids hallucination, and produces responses that are both accurate and grounded in the retrieved context. The MLLM processes the retrieved information from the previous phase and generates a contextually relevant, task-aligned response, which is then presented to the user.

4- Experiment Setup

4-1- Data Preprocessing

Although the original SLAKE dataset contains both English and Chinese question-answer pairs, this study focuses exclusively on the English subset to allow a clearer assessment of the specific impact of the MMKG structure within the multimodal RAG pipeline. Including Chinese data could introduce additional factors, such as cross-lingual alignment issues and language-specific retrieval behaviour, which may complicate the analysis and make it more difficult to isolate the effects of the MMKG itself. Therefore, a refined subset was extracted, consisting of 642 medical images and 7033 English question-answer pairs. The dataset was then randomly split at the image level using an 80% training and 20% testing split. As a result, the training set includes 513 images and 5639 question-answer pairs, which are used to construct the knowledge base for the multimodal RAG system. The test set comprises the remaining 129 images and 1394 question-answer pairs and is used to evaluate the performance of the proposed method.

4-2- Model Configuration

To evaluate the effectiveness of the proposed VisGraphRAG, this study compares it against a baseline MLLM and a baseline RAG. The baseline MLLM refers to a standard Multimodal Large Language Model (MLLM) that performs multimodal question answering directly without external knowledge retrieval, relying solely on its internal pre-trained parameters to generate answers based on the provided image-question pair. In this study, the MLLM used is Llama-4-Maverick-17B-128e.

On the other hand, the baseline RAG refers to a standard RAG setup that utilizes a vector database as its external knowledge source. The primary distinction between the two lies in their knowledge retrieval methods. VisGraphRAG utilizes a MMKG stored in a Neo4j graph database, retrieving information through Cypher queries and cosine similarity. In contrast, the baseline RAG system employs a conventional vector-based retrieval approach, using Qdrant as its database and relying solely on cosine similarity for retrieval. To ensure a fair comparison, both systems share the same core configuration wherever possible. Specifically, they both use CLIP-ViT-Large-Patch-14 for generating multimodal embeddings and Llama-4-Maverick-17B-128e as the response generation model. The top-k retrieval parameter is set to

3, and the temperature for generation is fixed at 0, ensuring deterministic outputs and reproducibility. Moreover, the maximum completion token limit is set to 512 to constrain the length of the generated responses and manage computational cost.

A summary of the configuration for both systems is provided in Table 1. This controlled comparison allows us to isolate the impact of structured knowledge representation through MMKGs on retrieval quality and generation accuracy in multimodal RAG systems.

Table 1. Model Configuration

Component	VisGraphRAG	Baseline RAG
Database	Graph Database: Neo4j	Vector Database: Qdrant
Vector similarity metric	Cypher + cosine similarity	Cosine similarity
Embedding Model	Clip-vit-large-patch14	
Generator Model	Llama-4-Maverick-17B-128e	
Top-k retrieval	3	
Temperature	0	
Max Completion Token	512	

4-3- Evaluation Metrics

To evaluate the performance of the proposed multimodal RAG system, this study adopts the Retrieval-Augmented Generation Assessment (RAGAS) framework introduced by Es et al. [24]. RAGAS evaluates three key aspects of response quality: faithfulness, answer relevance, and context relevance.

Faithfulness assesses whether the generated answer is grounded in the retrieved context, ensuring that each claim made can be directly inferred from the supporting evidence. Each generated answer is decomposed into individual statements, which are then verified by a Multimodal LLM against the context. The faithfulness score, F, is calculated as the proportion of supported statements. The final faithfulness score is calculated as Equation 2, where V is the number of statements supported by the context and S is the total number of statements.

$$F = \frac{V}{S} \tag{2}$$

Answer relevance evaluates how well the generated response addresses the original question. This is measured by using multimodal LLM to generate alternative questions from the answer and computing the cosine similarity between their embeddings and the original question. Higher similarity indicates stronger alignment with the intended query. The answer relevance score, AR, is computed as Equation 3, where sim $(q\ , q_i)$ is similarity score between the original question and the i-th generated question, and n is number of generated questions.

$$AR = \frac{1}{n}\sum_{i=1}^{n} sim\ (q, q_i) \tag{3}$$

Context relevance measures the conciseness and focus of the retrieved information. In this metric, a Multimodal LLM is used to extract key sentences necessary for answering the question. If no essential information is found from the retrieved context, the context is considered insufficient. Equation 4 shows the formula to calculate context relevance score, CR.

$$CR = \frac{number\ of\ extracted\ sentences}{total\ number\ of\ sentences\ in\ c(q)} \tag{4}$$

While the original RAGAS framework provides valuable insights into response quality by assessing faithfulness, answer relevance, and context relevance, it does not explicitly measure the factual accuracy of the generated answer or its alignment with multimodal inputs such as images. To address these limitations, this study incorporates three additional RAGAS metrics: Answer Accuracy, Multimodal Faithfulness, and Multimodal Relevance.

Answer Accuracy evaluates the degree to which the generated answer aligns with a predefined ground truth. This metric uses a dual “LLM-as-a-judge” setup, where a single LLM is prompted from two complementary perspectives to enhance robustness. In the first prompt, the LLM compares the generated answer to the ground truth. In the second, the roles are reversed, and the LLM compares the ground truth to the generated answer. Each comparison is rated on a scale of 0 (inaccurate), 2 (partially correct), or 4 (fully correct). These scores are then normalized to a [0,1] scale, and if both ratings are valid, their average is taken as the final Answer Accuracy score.

Multimodal Faithfulness assesses whether the generated answer is factually supported by either the textual or visual context. An answer is considered faithful only if all claims it contains can be directly inferred from the retrieved visual and/or textual evidence.

Multimodal Relevance measures how well the generated answer pertains to the retrieved information from both modalities. If the response aligns meaningfully with either the visual or textual inputs, it is deemed relevant. Similar to the faithfulness score.

All RAGAS evaluations are conducted using Gemini Flash 2.0, an innovative MLLM that serves as a unified evaluator across both textual and visual modalities. All metric scores range from 0 to 1, with higher values indicating better performance.

Since the RAGAS evaluation heavily relies on LLM-as-a-judge scoring, an additional human evaluation was conducted to assess the reliability and consistency of these automated judgments. A random sample of 50 questions was selected from the test set. Each generated answer was manually reviewed by a human annotator using the same scoring rubric as the one applied by the LLM-based evaluation. To quantify the level of agreement between the human and LLM judgments, inter-rater agreement was computed using percent agreement, PA. Percent agreement is the simplest and most intuitive form of inter-rater reliability. It is calculated by dividing the number of instances where both raters assigned the same score by the total number of evaluated items, then multiplying the result by 100 [25]. The formula is presented in Equation 5. However, among the six RAGAS metrics used in this study, only five were included in the human and LLM agreement analysis. Answer relevance was excluded because it depends on automatically generating synthetic questions and computing cosine similarity between vector embeddings. This approach cannot be consistently replicated or validated by human annotators, making it unsuitable for manual evaluation.

$$PA = \frac{number\ of\ agreements}{total\ number\ of\ judgements} \times 100 \tag{5}$$

To assess the efficiency of the retrieval mechanisms in both the baseline RAG and the proposed VisGraphRAG systems, we evaluate two timing-based metrics: retrieval time and total response time. Retrieval time refers to the duration taken to retrieve the top-k relevant nodes from the knowledge base, either vector database or MMKG. This timing begins immediately after the input question is encoded into a query embedding and ends once the relevant data has been retrieved. On the other hand, total response time measures the overall time required to generate a final answer, starting from the moment the input question is submitted and ending when the answer is fully produced. These measurements were conducted using 100 questions selected from the SLAKE dataset. All timings were recorded using standard Python profiling tools, and the minimum, maximum, and average values were calculated to compare system performance.

5- Results and Discussion

5-1- Results of RAGAS Evaluation

Table 2 and Figure 5 present the RAGAS evaluation scores for each model when tested on 20% of the SLAKE dataset. As the baseline Multimodal LLM does not perform external knowledge retrieval, most RAGAS metrics are not applicable to its evaluation. Therefore, it is assessed solely based on answer accuracy.

Table 2. Experimental Results for Baseline LLM, Baseline RAG and VisGraphRAG.

Model	Answer Accuracy	Faithfulness	Multimodal Faithfulness	Context Relevance	Multimodal Relevance	Answer Relevance
Llama-4-Maverick-17B-128e (Baseline MLLM)	0.5805	-	-	-	-	-
Baseline RAG	0.6743	0.6382	**0.9075**	0.7782	0.7912	**0.8264**
VisGraphRAG (Proposed Method)	**0.7629**	**0.8542**	0.8637	**0.9645**	**0.8802**	0.8252

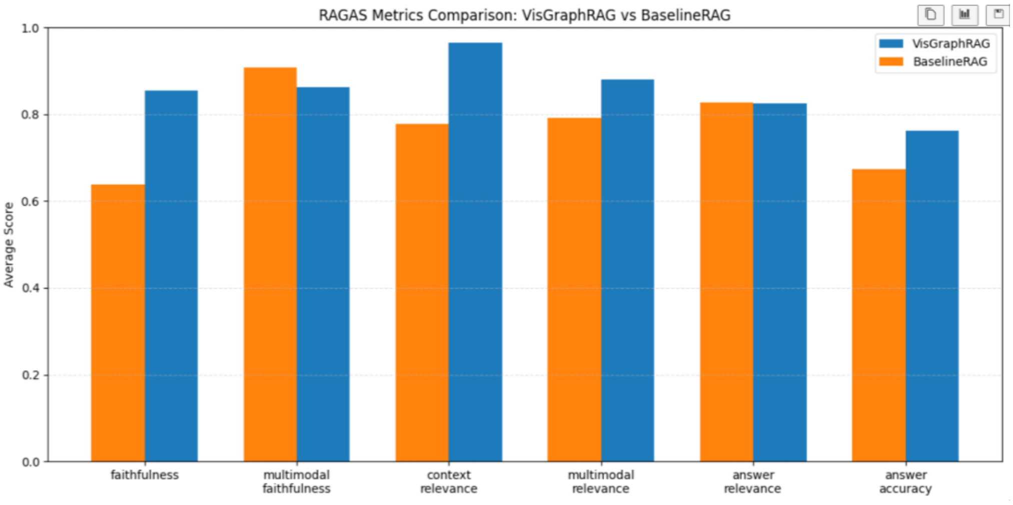


Figure 5. Bar Chart of RAGAS Metrics Comparison

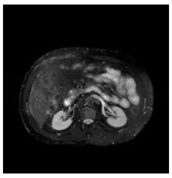
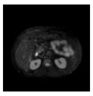
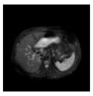
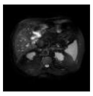
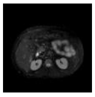
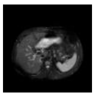
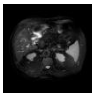
First and foremost, the most important metrics is the answer accuracy. Based on the results, the baseline MLLM achieved the lowest answer accuracy at 0.5805. The baseline RAG approach improved this to 0.6743, while the proposed VisGraphRAG method achieved the highest accuracy at 0.7629. These results show that incorporating RAG can significantly enhance the performance of a baseline MLLM. Furthermore, the RAG approach based on a multimodal knowledge graph demonstrates better performance compared to the vector database-based RAG.

In terms of faithfulness, VisGraphRAG achieves a score of 0.8542, significantly outperforming the baseline RAG, which scores 0.6382. This suggests that VisGraphRAG generates answers that are more accurately grounded in the retrieved textual context, reducing reliance on the implicit knowledge of the underlying multimodal language model (MLLM). However, when evaluated based on multimodal faithfulness, VisGraphRAG achieves a slightly lower score (0.8637) compared to the baseline RAG (0.9075). This slight reduction in multimodal faithfulness may be attributed to the fact that the retrieved text in VisGraphRAG is often semantically rich and contextually sufficient to answer the question accurately. Consequently, the MLLM may focus more on the rich textual context and engage less with the retrieved images during response generation. In contrast, the baseline RAG system often fails to retrieve the most relevant textual answers, which compels the model to rely more on the retrieved image content to compensate for the insufficient text. Nevertheless, this stronger visual grounding does not necessarily lead to better performance, as the answer accuracy of the baseline RAG remains lower. This outcome implies that a greater dependence on visual information but without effective cross-modal alignment may not contribute meaningfully to answer accuracy.

Moreover, VisGraphRAG also achieves a higher context relevance score of 0.9645, significantly outperforming the baseline RAG's score of 0.7782. This indicates that the text retrieved by VisGraphRAG is more pertinent and supportive in answering the question, highlighting the strength of MMKGs in retrieving semantically richer and task-relevant context. In addition, VisGraphRAG records a higher multimodal relevance score (0.8802) compared to the baseline RAG (0.7912), demonstrating its superior ability to retrieve relevant information across modalities. Unlike the baseline RAG, which treats images and questions as independent inputs, often resulting in contextual mismatches, VisGraphRAG models images, questions, and answers as interconnected entities within a unified graph structure. This integrated representation ensures that retrieved content maintains proper alignment between modalities, producing more coherent and contextually accurate answers.

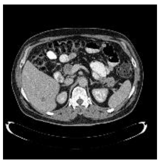
As illustrated in Table 3, the baseline RAG system occasionally produces incorrect answers because of misalignment between retrieved modalities. For instance, the correct answer to a query might be "MRI," yet the retrieved textual context describes "CT" instead. Even though the retrieved image is indeed an MRI scan, the model may misinterpret it as a CT image due to the mismatch in accompanying text. This confusion highlights a critical flaw in traditional vector-based RAG systems, which retrieve image and text embeddings independently. Such independent retrieval can cause cross-modal inconsistencies, where the retrieved text does not correspond accurately to the retrieved image. In contrast, VisGraphRAG overcomes this issue by structurally linking each image with its corresponding question and answer within a unified multimodal knowledge graph. This integrated approach ensures that the retrieved context remains semantically aligned across modalities, supporting more accurate and contextually grounded responses.

Table 3. Example of cross modalities misalignment

Query	Ground Truth	Model	Retrieved Context	Generated Answer
 What modality is used to take this image?	MRI	Baseline RAG	**Text:** 1. "**question**": "What modality is used to take this image?", "**answer**": "CT" 2. "**question**": "What modality is used to take this image?", "**answer**": "CT" 3. "**question**": "What modality is used to take this image?", "**answer**": "CT" **Image:** **1.**  **2.**  **3.** 	The modality used to take this image is CT.
		VisGraphRAG	 "**question**": "What modality is used to take this image?", "**answer**": "MRI"  "**question**": "What modality is used to take this image?", "**answer**": "MRI"  "**question**": "What modality is used to take this image?", "**answer**": "MRI"	The modality used to take this image is MRI.

Next, when examining the answer relevance metric, both VisGraphRAG and the baseline RAG achieve similar scores, with VisGraphRAG scoring 0.8252 and the baseline RAG slightly higher at 0.8264. This indicates that both models are generally capable of generating answers that align with the intent of the user query. Although VisGraphRAG's score is marginally lower, this difference is not particularly meaningful, as the evaluation mechanism for answer relevance may not always reflect actual correctness. This limitation is illustrated in Table 4, where both models generate responses to the same query. In this example, although the baseline RAG produces an incorrect answer ("The image looks normal") and VisGraphRAG provides the correct one ("The image does not look normal"), the answer relevance score is still higher for the baseline RAG. This discrepancy arises because RAGAS evaluates answer relevance by generating questions from the model's answer and then measuring its similarity to the original question. As a result, semantically incorrect answers can still receive high relevance scores if they share superficial lexical similarity with the query. This example highlights the limitations of automatic relevance scoring.

Table 4. An example of answer relevance scoring issue

Query	Ground Truth	Model	Generated Answer	Answer Relevance Score
 Does this image look normal?	No	VisGraphRAG	The image looks normal	0.7015
		VisGraphRAG	The image does not look normal.	0.6639

In summary, the results show that MMKG-based multimodal RAG systems, such as VisGraphRAG, significantly improve cross-modal alignment, enabling more accurate and contextually relevant answers compared to traditional vector-based RAG approaches. The structured nature of the MMKG allows the system to retrieve semantically coherent and task-relevant information by explicitly modelling the relationships between visual and textual entities. This, in turn, leads to responses that are not only more accurate but also better grounded in the retrieved context. These findings highlight the effectiveness of integrating structured multimodal knowledge into the RAG pipeline and demonstrate the potential of MMKG as a powerful external knowledge source for enhancing multimodal reasoning and generation.

5-2- Results of Human Evaluation

Table 5 presents the percentage agreement between human and LLM-based evaluations across five RAGAS metrics. The "Matched Questions" column indicates the number of instances where the human and LLM scores were identical, based on a sample of 50 questions. The results show that all metrics achieved consistently high levels of agreement, with values exceeding 90%. Although perfect alignment was not reached, the overall agreement remains strong, supporting the reliability and consistency of the LLM-as-a-judge scoring approach used in RAGAS.

Table 5. Percentage Agreement Between Human and LLM Evaluations

Metrics	Percent agreement (%)	Matched Questions (out of 50)
Answer Accuracy	90.00	45
Faithfulness	98.00	49
Multimodal Faithfulness	90.00	45
Context Relevance	96.00	48
Multimodal Relevance	92.00	46

5-3- Results of Retrieval and Response Speed Evaluation

Table 6 presents the benchmarking results for retrieval time and total response time across both models.

Table 6. Time Comparison Between VisGraphRAG and Baseline RAG

Model	Retrieval Time (sec)			Total Response Time (sec)		
	Min	Max	Avg	Min	Max	Avg
Baseline RAG	**0.426**	**2.667**	**0.676**	**0.629**	**4.545**	**2.459**
VisGraphRAG (Proposed Method)	1.973	8.742	3.687	5.074	18.926	8.711

The baseline RAG system achieves a significantly faster average retrieval time of 0.676 seconds and an average total response time of 2.459 seconds. This efficiency is attributed to its reliance on direct vector similarity search over pre-computed embeddings. In contrast, the proposed VisGraphRAG system records an average retrieval time of 3.687 seconds and a total response time of 8.711 seconds. The observed performance gap is expected, as the retrieval process involves additional computation such as Cypher-based graph traversal, relationship filtering, and cosine similarity ranking within the multimodal knowledge graph.

Despite the increased latency, this hybrid retrieval mechanism supports more semantically coherent and structurally aligned retrieval by maintaining explicit relationships between modalities. The trade-off in speed is considered acceptable given the benefit of improved retrieval quality and enhanced grounding in generated responses.

6- Conclusion

In conclusion, this study introduces VisGraphRAG, a novel multimodal RAG framework that leverages a Multimodal Knowledge Graph (MMKG) as an external database. Unlike traditional vector-based RAG systems, which retrieve information based on independent modality-specific embeddings, VisGraphRAG leverages the structured relationships within MMKG to enhance cross-modal alignment. By explicitly linking images, questions, and answers in a graph-based format, the proposed framework addresses the common issue of cross-modal misalignment and supports more coherent and contextually grounded generation. Experimental results demonstrate that VisGraphRAG consistently outperforms the baseline vector-based RAG model across multiple RAGAS metrics, particularly in faithfulness, context relevance, multimodal relevance and overall answer accuracy. These findings highlight the potential of graph-structured multimodal retrieval to support more reliable and interpretable multimodal reasoning in complex tasks.

Despite its promising results, this study has several limitations that present opportunities for future improvement. First, the current implementation generates text-only responses, as the LLaMA4 Vision model used in this work does not yet support true multimodal (image-text) output generation. Second, the evaluation was conducted on a relatively small and domain-specific test set, which, while adequate for initial validation, may limit the generalizability of the findings. Future work may involve extending VisGraphRAG to support multimodal output generation, allowing the model to produce responses that combine both textual and visual elements in a coherent manner. In addition, the MMKG can be expanded to cover a broader range of domains and incorporate a wider variety of cross-modal relationships, which would further enhance the system's reasoning capabilities. Moreover, evaluating the framework on larger, multilingual, and more complex datasets would provide a more comprehensive understanding of its scalability, robustness, and potential applicability in real-world multimodal scenarios.

7- Declarations

7-1- Author Contributions

Conceptualization, S.K.H. and L.Y.O.; methodology, S.K.H.; software, S.K.H.; validation, S.K.H., L.Y.O., and M.C.L.; formal analysis, S.K.H., L.Y.O., and M.C.L.; investigation, S.K.H.; resources, S.K.H., L.Y.O., and M.C.L.; data curation, S.K.H.; writing—original draft preparation, S.K.H.; writing—review and editing, S.K.H., L.Y.O., and M.C.L.; visualization, S.K.H.; supervision, L.Y.O. and M.C.L.; project administration, L.Y.O.; funding acquisition, L.Y.O. and M.C.L. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here [22].

7-3- Funding

This work was supported by the Telekom Malaysia Research and Development under Grant RDTC/221073 (MMUE/230002).

7-4- Institutional Review Board Statement

Not applicable.

7-5- Informed Consent Statement

Not applicable.

7-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. High-Confidence Computing, 4(2), 100211. doi:10.1016/j.hcc.2024.100211.

[2] Chen, Z., Xu, C., Qi, Y., Jiang, X., & Guo, J. (2025, November). VLM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training. In Findings of the Association for Computational Linguistics: EMNLP 2025, 8140-8158. doi:10.18653/v1/2025.findings-emnlp.432.

[3] Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2024). NExT-GPT: Any-to-Any Multimodal LLM. Proceedings of Machine Learning Research, 235, 53366–53397. doi:10.48550/arXiv.2309.05519.

[4] Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., Yan, H., Fu, J., Gui, T., Sun, T., Jiang, Y. G., & Qiu, X. (2024). AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, 9637–9662. doi:10.18653/v1/2024.acl-long.521.

[5] Singh, A. (2025). Meta Llama 4: The Future of Multimodal AI. SSRN Electronic Journal (Preprint), 1-15. doi:10.2139/ssrn.5208228.

[6] Zhao, R., Chen, H., Wang, W., Jiao, F., Do, X. L., Qin, C., ... & Joty, S. (2023). Retrieving multimodal information for augmented generation: A survey. arXiv Preprint, arXiv:2303.10868. doi:10.48550/arxiv.2303.10868.

[7] Chen, W., Hu, H., Chen, X., Verga, P., & Cohen, W. W. (2022). MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 5558–5570. doi:10.18653/v1/2022.emnlp-main.375.

[8] Riedler, M., & Langer, S. (2024). Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. arXiv Preprint, arXiv:2410.21943. doi:10.48550/arXiv.2410.21943.

[9] Joshi, P., Gupta, A., Kumar, P., & Sisodia, M. (2024). Robust Multi Model RAG Pipeline for Documents Containing Text, Table & Images. Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2024, 993–999. doi:10.1109/ICAAIC60222.2024.10574972.

[10] Zheng, J., Liang, M., Yu, Y., Li, Y., & Xue, Z. (2024). Knowledge Graph Enhanced Multimodal Transformer for Image-Text Retrieval. Proceedings - International Conference on Data Engineering, 70–82. doi:10.1109/ICDE60146.2024.00013.

[11] Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., & Yao, H. (2025). Mmed-Rag: Versatile Multimodal Rag System for Medical Vision Language Models. In 13th International Conference on Learning Representations, ICLR 2025, 76330–76359. doi:10.48550/arxiv.2410.13085.

[12] Mankari, S., & Sanghavi, A. (2024). Enhancing Vector based Retrieval Augmented Generation with Contextual Knowledge Graph Construction. 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry, IDICAIEI 2024, 1–6. doi:10.1109/IDICAIEI61867.2024.10842699.

[13] Shavaki, M. A., Omrani, P., Toosi, R., & Akhaee, M. A. (2024). Knowledge Graph Based Retrieval-Augmented Generation for Multi-Hop Question Answering Enhancement. 15th International Conference on Information and Knowledge Technology, IKT 2024, 78–84. doi:10.1109/IKT65497.2024.10892619.

[14] Lee, J., Wang, Y., Li, J., & Zhang, M. (2024). Multimodal Reasoning with Multimodal Knowledge Graph. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, 10767–10782. doi:10.18653/v1/2024.acl-long.579.

[15] Chen, Z., Zhang, Y., Fang, Y., Geng, Y., Guo, L., Chen, X., Li, Q., Zhang, W., Chen, J., Zhu, Y., Li, J., Liu, X., Pan, J. Z., Zhang, N., & Chen, H. (2024). Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. arXiv Preprint, arXiv: 2402.05391. doi:10.48550/arXiv.2402.05391.

[16] Yu, J., Zhu, Z., Wang, Y., Zhang, W., Hu, Y., & Tan, J. (2020). Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recognition, 108, 107563. doi:10.1016/j.patcog.2020.107563.

[17] Zheng, J., Liang, M., Yu, Y., Du, J., & Xue, Z. (2024). Multimodal Knowledge Graph-Guided Cross-Modal Graph Network for Image-Text Retrieval. Proceedings - 2024 IEEE International Conference on Big Data and Smart Computing, BigComp 2024, 97–100. doi:10.1109/BigComp60711.2024.00024.

[18] Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). MM-LLMs: Recent advances in multimodal large language models. arXiv Preprint, arXiv:2401.13601. doi:10.48550/arXiv.2401.13601.

[19] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 43(2), 3703155. doi:10.1145/3703155.

[20] Kau, A., He, X., Nambissan, A., Astudillo, A., Yin, H., & Aryani, A. (2024). Combining knowledge graphs and large language models. arXiv preprint, arXiv:2407.06564. doi:10.48550/arxiv.2407.06564.

[21] Chen, Y., Ge, X., Yang, S., Hu, L., Li, J., & Zhang, J. (2023). A survey on Multimodal knowledge Graphs: construction, completion and applications. Mathematics, 11(8), 1815. doi:10.3390/math1108181.

[22] Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., & Wu, X. M. (2021). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. Proceedings - International Symposium on Biomedical Imaging, 2021-April, 1650–1654. doi:10.1109/ISBI48211.2021.9434010.

[23] Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. Proceedings of 2016 4th International Conference on Cyber and IT Service Management, CITSM 2016. doi:10.1109/CITSM.2016.7577578.

[24] Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150-158. doi:10.18653/v1/2024.eacl-demo.16.

[25] Halpin, S. N. (2024). Inter-Coder Agreement in Qualitative Coding: Considerations for its Use. American Journal of Qualitative Research, 8(3), 23–43. doi:10.29333/ajqr/14887.