IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 March 2024, accepted 14 May 2024, date of publication 20 May 2024, date of current version 28 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3403101

RESEARCH ARTICLE

UniRaG: Unification, Retrieval, and Generation for Multimodal Question Answering With Pre-Trained Language Models

QI ZHI LIM[1], CHIN POO LEE[1], (Senior Member, IEEE), KIAN MING LIM[1], (Senior Member, IEEE), AND AHMAD KAMSANI SAMINGAN[2]

[1]Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia
[2]Telekom Research and Development Sdn. Bhd., Cyberjaya, Selangor 63000, Malaysia

Corresponding author: Chin Poo Lee (cplee@mmu.edu.my)

This work was supported by the Telekom Malaysia Research and Development Grant under Award RDTC/231075 and Award RDTC/231084.

ABSTRACT Multimodal Question Answering (MMQA) has emerged as a challenging frontier at the intersection of natural language processing (NLP) and computer vision, demanding the integration of diverse modalities for effective comprehension and response. While pre-trained language models (PLMs) exhibit impressive performance across a range of NLP tasks, the investigation of text-based approaches to address MMQA represents a compelling and promising avenue for further research and advancement in the field. Although recent research has delved into text-based approaches for MMQA, the attained results have been unsatisfactory, which could be attributed to potential information loss during the knowledge transformation processes. In response, a novel three-stage framework named UniRaG is proposed for tackling MMQA, which encompasses unified knowledge representation, context retrieval, and answer generation. At the initial stage, advanced techniques are employed for unified knowledge representation, including LLaVA for image captioning and table linearization for tabular data, facilitating seamless integration of visual and tabular information into textual representation. For context retrieval, a cross-encoder trained on sequence classification is utilized to predict relevance scores for question-document pairs, and a top-k retrieval strategy is employed to retrieve the documents with the highest relevance scores as the contexts for answer generation. Finally, the answer generation stage is facilitated by a text-to-text PLM, Flan-T5-Base, which follows the encoder-decoder architecture with attention mechanisms. During this stage, uniform prefix conditioning is applied to the input text for enhanced adaptability and generalizability. Moreover, contextual diversity training is introduced to improve model robustness by including distractor documents as negative contexts during training. Experimental results on the MultimodalQA dataset demonstrate the superior performance of UniRaG, surpassing the existing state-of-the-art methods across all scenarios with 67.4% EM and 71.3% F_1. Overall, UniRaG showcases robustness and reliability in MMQA, heralding significant advancements in multimodal comprehension and question answering research.

INDEX TERMS Computer vision, information retrieval, multimodal question answering, natural language processing, pre-trained language models, unified knowledge representation.

I. INTRODUCTION

In response to the increasing demand for more sophisticated and contextually aware artificial intelligence applications, Multimodal Question Answering (MMQA) has emerged as a revolutionary paradigm, strategically designed to address the limitations of traditional question answering (QA) approaches. MMQA expands the scope and capabilities of QA systems by seamlessly integrating information from various modalities, including text, tables, and images [1],

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

[2]. As the digital landscape becomes more complex, MMQA presents itself as a pivotal solution that extends beyond linguistic comprehension, leveraging the information from multiple modalities to enhance the understanding and responsiveness of QA systems. In the context of MMQA, textual information remains a vital component, providing the foundation for linguistic comprehension. Additionally, the integration of tables enables the extraction of structured data, enhancing the system's ability to process and analyze tabular information. Furthermore, the inclusion of images introduces a visual dimension to the system, allowing it to interpret and respond to queries based on visual content.

Previously, research efforts within the QA domain predominantly concentrated on unimodal approaches, with a specific emphasis on text-centric, table-based, and visual question answering. Text QA systems focused on extracting information and generating responses solely from textual data [3], [4], [5], while table QA methodologies delved into structured datasets, extracting insights primarily from tabular formats [6], [7], [8]. On the other hand, visual question answering aims at deciphering and responding to queries based on visual content [9], [10], [11]. While these unimodal approaches have made significant advancements, they have inherently overlooked the potential synergy that arises from the simultaneous integration of multiple modalities, including text, tables, and images. The evolving landscape of artificial intelligence (AI) research now compels a shift towards embracing the richness and complexity offered by multimodal frameworks. This paradigm shift seeks to bridge the gaps left by unimodal systems, enabling a more comprehensive understanding of information by leveraging the collective power of diverse data modalities.

However, MMQA presents formidable challenges due to the intricate nature of handling multiple modalities [1], [2]. Firstly, the process involves retrieving the relevant supporting documents from a curated list of candidate documents, thereby adding layers of complexity to the task. This preliminary step ensures that the model accesses a diverse set of information sources, optimizing its ability to respond accurately to a wide range of queries. Subsequently, a robust QA model is required to predict answers to the questions based on the multimodal contexts retrieved. In this case, the complexity of multi-hop reasoning necessitates the development of advanced algorithms and strategies to effectively navigate and extract meaningful insights from the diverse information sources encompassed by MMQA. Moreover, the integration of textual, tabular, and visual data further complicates the MMQA process, demanding sophisticated techniques for data fusion and cross-modal understanding to achieve comprehensive and accurate question answering capabilities.

In addressing the challenges of MMQA, three primary approaches have been explored by researchers in recent research. The first involves training separate QA models for different modalities and decomposing the questions into several sub-questions for step-by-step reasoning [1], [2], [12]. Although this approach is simple and straightforward for solving MMQA, it may encounter difficulties when answering questions that require cross-modal reasoning as there is no established interaction between the models used. Additionally, information loss may occur during the step-by-step reasoning process, leading to suboptimal question answering performance. The second approach advocates for the development of a single, multimodal model capable of processing inputs from diverse sources and modalities simultaneously [13], [14], [15], thereby generating the final answer to the question. By incorporating strategies such as vision-language pre-training, the multimodal model can have a more nuanced understanding of multimodal contexts, enhancing its capability for cross-modal reasoning and facilitating enhanced question answering performance. However, in order to accommodate the variety of inputs, this approach requires extensive pre-training and fine-tuning, which leads to the practical challenge of demanding substantial computational resources and potential issues related to overfitting.

The third strategy employed in tackling the challenge of MMQA adopts a text-based approach, providing an alternative solution to the complexities associated with multimodal reasoning [16], [17], [18]. In this approach, the first step involves the conversion of multimodal knowledge into a unified textual representation. By transforming diverse modalities into a common language, this method seeks to overcome the difficulties arising from the disparate nature of modalities, ensuring a cohesive and standardized input for subsequent processing. Following the unification process, the retrieval and question answering tasks are executed using a text-based approach. This entails leveraging advanced natural language processing (NLP) techniques and models designed for textual comprehension to navigate through the unified representation and derive meaningful responses. Unlike the first approach, which relies on separate models for individual modalities, and the second approach, which demands extensive pre-training for a single multimodal model, this text-based strategy aims to capitalize on the inherent strengths of textual understanding while potentially mitigating challenges associated with information loss and computational resource demands.

Furthermore, the third strategy harnesses the power of pre-trained language models (PLMs) to bolster their effectiveness in handling MMQA. By leveraging the capabilities of state-of-the-art language models pre-trained on vast amounts of diverse textual data [19], [20], [21], this approach capitalizes on a nuanced understanding of language, context, and semantics. PLMs serve as a robust foundation for solving MMQA using a text-based strategy, enabling effective interpretation and processing of the unified textual representations derived from multimodal data. The strength of PLMs lies in their ability to capture intricate patterns, contextual relationships, and domain-specific knowledge, thereby enhancing the system's

comprehension of diverse contexts. The utilization of PLMs not only facilitates more accurate question answering but also contributes to the adaptability of the system across various MMQA scenarios. In other words, the strategic integration of PLMs aligns with contemporary advancements in NLP and signifies a promising avenue for addressing the intricate challenges posed by MMQA.

Therefore, in this study, the text-based approach outlined as the third strategy is strategically adopted to address MMQA challenges [2]. By mapping multimodal knowledge into a unified textual representation, the strength of PLMs [19], [20] is leveraged for enhanced multimodal retrieval and question answering performance. Specifically, a novel three-stage framework named UniRaG is proposed for solving MMQA, encompassing unified knowledge representation, context retrieval, and answer generation. At the initial stage, the multimodal knowledge is transformed into a unified textual representation through advanced image-to-text and table-to-text techniques. Notably, LLaVA [22] is utilized for image captioning to generate comprehensive descriptions of images, which include the distinct features and objects present in the visual context. On the other hand, the table data is converted into textual representation through the table linearization technique [23], which effectively preserves the tabular structure with no information loss. With the multimodal knowledge represented in text, the subsequent retrieval and question answering processes can be strategically solved by incorporating text-based approaches.

For context retrieval, a cross-encoder trained on sequence classification [24] is used to predict relevance scores for each question-document pair. Specifically, the model employed is the state-of-the-art cross-encoder [24] pre-trained on the MS Marco Passage Ranking task [25], ms-marco-MiniLM-L-12-v2. It is derived from a distilled version of the BERT-Base model with 12 hidden layers and a hidden size of 384 [26]. Meanwhile, a top-k retrieval strategy is adopted with a specific value set to 3, selecting the top-3 documents with the highest relevance scores from the list of candidate documents as the contexts for answer generation. During the answer generation stage, the question and retrieved contexts are concatenated as input text and passed into a text-to-text PLM for generating the final answer to the question. Particularly, the model used is the large-scale instruction fine-tuned version of the T5 model [20], Flan-T5-Base [27], which has shown its superiority and robustness across various NLP tasks. Additionally, this research employs uniform prefix conditioning on the input text to enhance answer generation, providing semantic guidance to the model and improving its adaptability to various queries for more nuanced responses. Moreover, contextual diversity training is introduced in this study to enhance the robustness and generalization of the model by including distractor documents alongside true supporting documents during training. Exposing the model to a variety of contexts enhances its ability to differentiate between relevant information and noise, thereby reducing the risks of overfitting and improving the model's resilience and adaptability.

To validate the effectiveness of the proposed UniRaG framework, comprehensive experiments were conducted on the commonly used MMQA dataset, MultimodalQA [2]. The experimental results demonstrate that UniRaG has significantly outperformed all existing methods, showcasing state-of-the-art performance across all scenarios of the MultimodalQA dataset. Furthermore, rigorous ablation studies and experiments underscore the robustness and reliability of the proposed framework and methodologies in addressing the challenges of MMQA. In essence, the primary contributions of this research are as follows:

- Proposing a novel three-stage framework to effectively tackle MMQA, namely UniRaG. The proposed framework encompasses unified knowledge representation, context retrieval, and answer generation.
- Employing LLaVA to generate rich and detailed image descriptions, reducing information loss during image-to-text transformation. Additionally, utilizing table linearization to convert tabular data into textual representation, preserving the tabular structure with no information loss during table-to-text transformation.
- Utilizing a cross-encoder trained on sequence classification to predict relevance scores for question-document pairs and adopting a top-k retrieval strategy to retrieve the most relevant documents as contexts.
- Fine-tuning Flan-T5-Base, a generative pre-trained language model, for answer generation. Applying uniform prefix conditioning to provide semantic guidance, enabling more contextually nuanced responses.
- Introducing contextual diversity training, incorporating distractor or negative documents as contexts during training. This diversification enhances the model's ability to discern pertinent information, reducing overfitting and improving robustness and flexibility.

II. RELATED WORKS

In the dynamic field of QA research, the progression from unimodal to multimodal paradigms represents a significant evolution. The inception of QA systems was characterized by unimodal models, primarily tailored to process and respond to textual data using NLP techniques [4], [5]. As the demand for a more holistic understanding of information grew, the QA landscape expanded to include bimodal approaches, wherein the fusion of text with images or tables sought to enrich the contextual understanding of queries. Visual Question Answering (VQA) [10] emerged as the pioneer in this domain, which focused on answering questions based on visual inputs. Subsequent advancements in this domain, such as OK-VQA [28] and KVQA [29], have extended the coverage of VQA by introducing questions that require knowledge from both image and textual data for accurate responses. On the other hand, researchers also delved into hybrid QA tasks that necessitate complex reasoning over tabular and textual data, such as HybridQA [30], OTT-

QA [31], and TAT-QA [32]. This leads to the development of more sophisticated frameworks that integrate both textual and visual or tabular elements in the context of question answering.

Building upon this trajectory, Multimodal Question Answering (MMQA) has emerged as a new focal point in the QA domain. Several MMQA datasets have been introduced [1], [2], [12], [33], [34], necessitating information from multiple modalities to effectively address the intricacies of question answering. Among these datasets, MultimodalQA [2] stands out as the quintessential, demanding proficiency in retrieval and question answering across a diverse spectrum of modalities, including textual content, tabular data, and images. In order to tackle the challenge of MMQA, it is observed that three main strategies have been adopted by researchers in recent studies. The first strategy entails training separate QA models for different modalities and breaking down questions into sub-questions for step-by-step reasoning [1], [2], [12]. The second approach facilitates a single multimodal model capable of simultaneously processing inputs from different sources and modalities to solve MMQA [13], [14], [15]. In this research, the proposed framework is meticulously devised based on the third strategy, which involves converting information from different modalities into a unified textual representation and incorporating text-based approaches to solve MMQA.

While PLMs demonstrate outstanding performance in various NLP tasks, exploring text-based approaches for tackling MMQA presents a compelling and promising avenue for further research and advancement. Numerous prior studies have actively engaged in exploring text-based approaches for MMQA [16], [17], [18], contributing valuable insights and paving the way for continued research in this domain. To effectively leverage text-based approaches for MMQA, a crucial step involves unifying multimodal data into textual representations. This process includes generating text descriptions for image data using computer vision techniques such as image captioning [35], [36] or object detection [37], [38], as well as transforming tabular data to text through methods like table linearization [23] or template-based [39] approaches. However, the process of converting images to text may entail the risk of information loss, potentially compromising the richness and detail inherent in visual data. In response, the research employed LLaVA [22] for image captioning, a framework capable of generating comprehensive and detailed image descriptions, thereby mitigating the impact of information loss during the transformation process. On the other hand, table linearization [23] is employed for table-to-text transformation, which retains all the information with the tabular structure. By consolidating multimodal data into the text space, text-based methods can be applied for the subsequent processes, fostering a more integrated and coherent analysis of diverse data modalities.

Information retrieval is the systematic process of searching for relevant information from a large corpus of data. In the context of question answering, information retrieval plays a pivotal role in extracting pertinent documents or passages containing answers to queries. Traditional information retrieval methods such as Term Frequency-Inverse Document Frequency (TF-IDF) [40] and BM25 (Best Matching 25) [41] have served as fundamental approaches, using statistical measures to rank documents based on their relevance to query terms. Recent advances in NLP have introduced more sophisticated techniques for effective information retrieval, particularly neural network-based models such as BERT (Bidirectional Encoder Representations from Transformers) [19]. These models utilize contextual embeddings to better comprehend the meaning of queries and documents, thereby enhancing the precision of information retrieval and subsequently improving the performance of question answering. In the current landscape of information retrieval, two primary methods have emerged as prominent strategies for enhancing the precision and efficacy of retrieving pertinent information. The first method involves encoding queries and documents using distinct encoders, followed by computing the similarity between the encoded representations [42], [43], [44]. This approach encompasses various techniques such as vector space models, word embeddings, and sentence/document embeddings, each offering nuanced ways to capture semantic relationships and context between queries and documents. The second method revolves around training a single model to predict the relevance score of query-document pairs directly [45], [46], [47]. This approach leverages supervised learning techniques, utilizing models like ranking models and Learning to Rank (LTR) algorithms to optimize ranking performance based on relevance labels in training data. While the first approach offers an in-depth understanding of the queries and documents, the second method provides direct learning of relevance between them, potentially leading to better generalization and improved performance in information retrieval tasks.

The QA task has undergone a notable evolution, transitioning from extractive methods, which directly retrieve answers from a given context, to generative approaches, where answers are synthesized based on comprehension and reasoning abilities. Extractive QA methods [19], [48], [49] typically involve identifying relevant snippets of text containing the answer to a given question, often utilizing techniques such as passage ranking and answer span prediction to get the final answer. However, these approaches may be constrained in their capacity to generate contextually appropriate responses. In contrast, generative QA models leverage language generation techniques to dynamically generate answers based on the understanding of the question and contexts. Recent advancements in generative QA have been catalyzed by the development of large-scale PLMs such as T5 (Text-to-Text Transfer Transformer) [20] and GPT (Generative Pre-trained Transformer) [21]. These models excel in understanding and generating natural language, facilitating more contextually relevant responses in question answering tasks. Furthermore, techniques like fine-tuning and multi-task learning have augmented the performance and

adaptability of generative QA systems, heralding a new era of sophisticated and effective question answering capabilities.

III. METHODOLOGY

MMQA can be defined as a task that involves retrieval and question answering in a multimodal context. In this task, a question Q is presented along with a set of candidate documents $D = \{d_0, d_1, \ldots, d_n\}$, where n represents the number of candidate documents. These documents can be presented in diverse formats, including text, tables, or images. The key challenges of MMQA revolve around accurately retrieving the relevant documents and generating precise answers based on the synthesized information from the documents retrieved. In this study, a novel three-stage framework named UniRaG is proposed, which delineates MMQA into three primary stages: unified knowledge representation, context retrieval, and answer generation.

Figure 1 depicts the overall framework of UniRaG. At the initial stage, multimodal knowledge in different document types is mapped into a unified textual representation. This process involves generating detailed image descriptions using an advanced image-to-text approach and transforming tables into natural language text through a designated table-to-text transformation. Following this, a context retrieval module is employed to predict the relevance scores for question-document pairs and retrieve the top-k documents from the list of candidate documents. These selected documents will serve as contextual information for the subsequent question answering process. Finally, the question and retrieved context are formatted into input text and passed into a generative PLM to generate the final answer to the given question.

A. UNIFIED KNOWLEDGE REPRESENTATION

In this study, text-based approaches are employed to address the challenges associated with MMQA. Preprocessing steps play a pivotal role in transforming knowledge represented in different modalities into a unified textual representation, which is essential for subsequent retrieval and question answering processes. Notably, the conversion of images and tables into textual representation stands out as a key aspect of this initial stage.

1) IMAGE-TO-TEXT

The transformation of images into natural language text presents a significant challenge due to the potential loss of information. Traditional image captioning models [35], [36] predominantly focus on generating concise and succinct image captions that only include a small fraction of the intricate details present in the visual content. Unfortunately, this brevity of the traditional models often results in the omission of crucial information, such as specific features and objects within the image. The limited scope of the generated captions impedes their ability to provide a comprehensive representation of the visual context, making them ill-suited for tasks that require a detailed understanding of the visual context, including information retrieval and question answering.

To mitigate information loss during conversion, this study utilizes the Large Language and Vision Assistant (LLaVA) to generate rich and detailed image descriptions. LLaVA [22], [50], an open-source Large Multimodal Model (LMM), was trained through fine-tuning LLaMA/Vicuna [51] on multimodal instruction-following data generated by GPT, a process known as visual instruction tuning. LLaVA comprises three main components: a vision encoder (CLIP), a vision-language connector (MLP), and a language model (Vicuna v1.5). Leveraging the advantages of a large language model with visual instruction tuning, LLaVA is capable of producing rich and detailed image descriptions that include specific features and objects within the visual context.

More specifically, the model used for image-to-text transformation is LLaVA-v1.5-7B [22]. In this study, a specific set of parameters for generation has been applied to enhance the quality and diversity of the generated content. First, the maximum number of output tokens is set to 512, which means that the generated image description will have a maximum length of 512 tokens. Besides, sample decoding is enabled with a temperature value of 0.2 and a top_p value of 0.7. The temperature parameter is set to a lower value to control the level of randomness in the sampling process, leading to more focused and deterministic text generation. On the other hand, the top_p parameter is set to 0.7 to strike a balance, allowing for a reasonable level of exploration and diversity while maintaining a degree of predictability and coherence in the generated content.

2) TABLE-TO-TEXT

The table-to-text transformation process is relatively simple and straightforward. No specific model is required for the conversion, as the table data is already presented in text. In this study, the table information is systematically formatted into a text representation that mirrors its tabular structure, which is commonly known as table linearization [23]. Specifically, in this study, all the information in the cell of the same row is concatenated into a single line and delimited by a vertical bar symbol ''|''. Besides, each row is assigned a unique identifier, denoted as ''row-id'', specifying the row number of the line of table information. Furthermore, ''-'' is inserted into the table cell with no information provided, representing that no information is applicable for the specific cell. In this case, the table data is fully transformed into text representations, while preserving the tabular structure and avoiding any loss of information during the transformation process.

B. CONTEXT RETRIEVAL

Since the MMQA dataset provides a list of candidate documents for each question, the context retrieval process plays a critical role in retrieving the supporting documents that provide answers to the questions. This step is important to

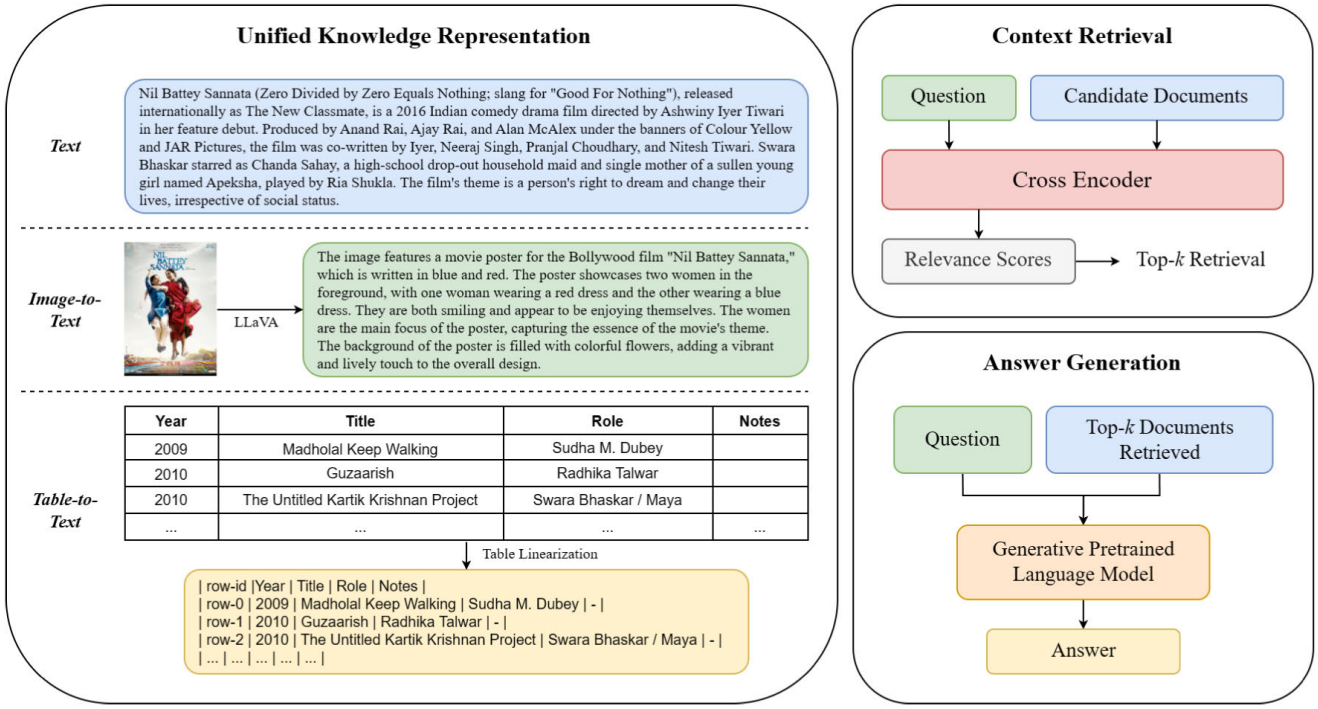

FIGURE 1. Overall framework of UniRaG.

filter out most of the distractors or negative samples from the list of candidate documents, and the final retrieved documents will serve as the contexts for answer generation. In this study, the strategy used for context retrieval is to select the top-k documents that are most relevant to each question from the candidate document list.

After unifying all the documents from different modalities into text representations, a cross-encoder based on sequence classification [24] is trained for context retrieval. Each question and document in the candidate document list are concatenated and input into the model to predict the relevance score for each question-document pair. Specifically, a BERT-based sequence classification model with a linear activation function is utilized in this study. The relevance score of a question-document pair can be calculated as (1):

$$s = \text{Linear}(\text{BERT}(Q; d)) \tag{1}$$

where Q is a question, d is a candidate document of the question, and s is the predicted relevance score to the question-document pair. This score indicates the relevance of the document to the given question as well as the necessity of that document to answer the question. By predicting the relevance scores of the question and all the candidate documents, the top-k documents with the highest relevance scores are selected as the context for answer generation.

C. ANSWER GENERATION

After retrieving the top-k documents for the question, a generative PLM is fine-tuned for answer generation. The selected model is T5 (Text-to-Text Transfer Transformer) [20], a transformer-based language model with an encoder-decoder architecture. T5's pre-training on large text datasets facilitates effective transfer learning, capturing general language patterns that are valuable for understanding and generating answers across diverse contexts. Additionally, T5's sequence-to-sequence architecture and attention mechanisms are crucial for handling variable input lengths and focusing on relevant information during answer generation. These attributes are particularly essential for MMQA, which include questions that require multiple contexts from different modalities to generate precise answers. Furthermore, the large-scale instruction fine-tuned version of the T5 model, Flan-T5-Base [27], is employed, which has shown improved performance in various NLP tasks such as reasoning and question answering. Figure 2 shows the detailed answer generation process and the architecture of the T5 model used in this study.

As shown in Figure 2(a), the question and the top-k documents retrieved are concatenated as input text, which is then passed into the T5 model for generating the final answer to the question. Notably, uniform prefix conditioning is applied to the input text to further enhance the performance of answer generation. The utilization of a standardized prefix serves as a form of semantic scaffolding for the model, guiding it to focus on critical elements or information within the concatenated input. This uniform prefix conditioning helps the model develop a more generalized understanding of the relationships between questions and relevant documents, fostering adaptability across diverse queries and facilitating a more contextually informed response.

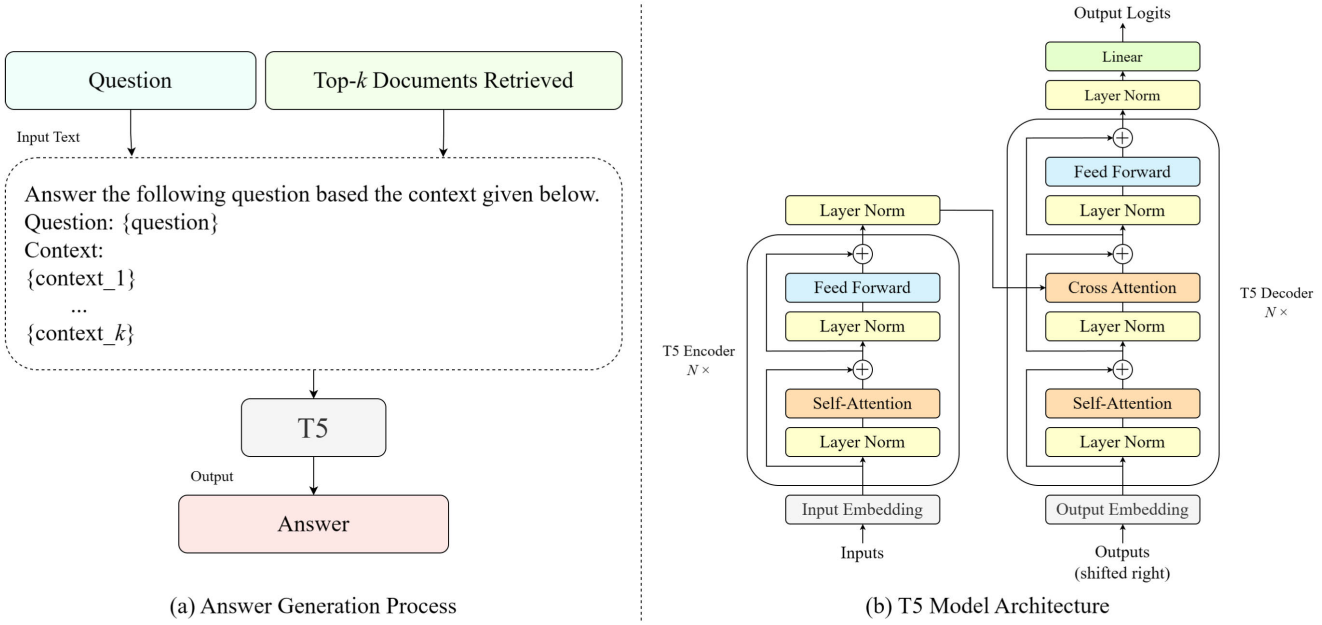

FIGURE 2. **(a) Answer generation process. (b) T5 model structure.**

Figure 2(b) illustrates the intricate model structure of the T5 model utilized in this study, which employs an encoder-decoder architecture with attention mechanisms [52]. Before inputting data into the encoder blocks, embeddings are initially generated through an embedding layer. These embeddings capture semantic information about the input tokens, facilitating subsequent analysis and manipulation within the transformer-based model. Conversely, prior to entering the decoder block, the expected outputs undergo preprocessing, wherein they are shifted right by one position and converted into embeddings. This preprocessing ensures that during training, the decoder receives the correct target input at each decoding step, aligning with the autoregressive nature of sequence generation tasks.

Unlike the original Transformer architecture [52], the T5 model eschews traditional positional encoding for providing information about the position of each token in the input sequence. Instead, T5 employs an innovative method known as relative position representations (RPR) to encode positional information during self-attention computation. RPR allows the model to capture positional information relative to other tokens in the sequence without explicitly encoding absolute positions, as in traditional positional encodings. This approach proves more efficient and effective for tasks where absolute positional information is less important.

The encoder of the T5 model comprises a stack of $N = 12$ identical blocks. Within each encoder block, layer normalization (Layer Norm) is strategically applied before every layer, including the self-attention mechanism and the subsequent feed-forward network. This normalization step stabilizes training by standardizing the activations of each layer. Subsequently, the model employs a self-attention mechanism to determine the relative importance of different words in the input sequence, computing attention scores for each word to generate context-aware representations. The output of the attention mechanism is then processed by the feed-forward network, capturing intricate data patterns through linear transformations interspersed with non-linear activation functions. Additionally, a residual connection is applied after each layer in the encoder block to facilitate gradient flow during training and mitigate the vanishing gradient problem. Finally, after completing the stack of $N = 12$ blocks in the encoder, a final layer normalization step is applied to the output, ensuring standardized representations before passing them to subsequent layers or tasks, maintaining stability and consistency in the model's architecture.

The decoder of the T5 model also comprises a stack of $N = 12$ identical blocks, with layer normalization applied before each layer for consistent training. Its self-attention mechanism operates similarly to that in the encoder block but focuses on previously generated tokens in the output sequence, capturing interdependencies within the target sequence. Additionally, the decoder block introduces an extra cross-attention mechanism, leveraging information from the encoder's output representations. This mechanism computes attention scores between the decoder's current hidden state and the encoder's output, providing context-aware representations for each position in the decoder relative to the entire input sequence. Subsequently, a feed-forward network processes the output of the attention mechanisms independently for each decoder position, capturing complex patterns and relationships within the data through linear transformations followed by non-linear activation functions. Residual connections between layers in the decoder blocks

ensure efficient training and improved model performance. Finally, a final layer normalization is applied to the decoder's output after completing the decoder operations.

After completing the operations of both encoder and decoder blocks in the T5 model, a single linear layer is applied to the output representations before proceeding to further processing or decoding. This linear layer serves as a final transformation step, mapping the high-dimensional representations learned by the model to the output space required for text generation. By applying the linear layer once after the encoder-decoder processing, the model seamlessly integrates the accumulated contextual information to generate coherent and contextually relevant text outputs. This unified approach ensures that the text generated by the model aligns with the intended task, providing accurate responses across various natural language generation tasks. In this study, the specific task is to generate precise answers to questions based on the provided contexts. Algorithm 1 summarizes the proposed UniRaG framework for MMQA.

Algorithm 1 Algorithm of the Proposed UniRaG Framework

Input: Question, Multimodal knowledge in different document types (text, images, tables).
Output: Final answer to the question.
Procedure:

1: **Unified Knowledge Representation:**
   a. Convert images into text descriptions using LLaVA.
   b. Transform tables into text through table linearization.
2: **Context Retrieval:**
   a. Predict relevance scores for question-document pairs using the chosen cross-encoder trained on sequence classification, ms-marco-MiniLM-L-12-v2.
   b. Retrieve top-k relevant documents as contexts.
3: **Answer Generation:**
   a. Concatenate question and retrieved contexts into input text with uniform prefix conditioning.
   b. Generate final answer to the question with the selected generative PLM, Flan-T5-Base.

D. MODEL TRAINING

In this research, the training process consists of two distinct phases: retrieval training and generation training. During retrieval training, a cross-encoder based on sequence classification is trained to predict relevance scores for question-document pairs. Subsequently, in the generation training phase, a generative PLM is fine-tuned to produce accurate answers based on the retrieved contexts.

1) RETRIEVAL TRAINING

For training the retrieval model, a systematic binary classification method is utilized. Each question is concatenated with every document in the candidate document list to create distinct question-document pairs for the training dataset. These pairs are then assigned binary labels, with 1 indicating a true supporting document and 0 representing a distractor or negative sample. This approach enables the model to learn the discriminative patterns necessary for accurately distinguishing between supporting documents and distractors.

The objective function employed for retrieval training is the Binary Cross Entropy with Logits Loss, which combines the Sigmoid activation function and binary cross-entropy loss. This loss function ensures stable numerical computation during the training process. Mathematically, the retrieval training loss function can be expressed as (2):

$$\mathcal{L}_{retr} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log\left(\sigma\left(x_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1\sigma\left(x_i\right)\right)\right) \tag{2}$$

where N is the number of samples in a batch, i is the index representing the i-th sample in the batch, y represents the ground truth label, x represents the output logits from the model, and σ denotes the Sigmoid activation function applied to transform the logits into probabilities.

2) GENERATION TRAINING

To train the answer generation model, each question and its supporting documents are concatenated as input text with uniform prefix conditioning. Since this study employs top-k retrieval, contextual diversity training is introduced to enhance the model's robustness and generalization. Unlike conventional approaches that only use true supporting documents for training, this novel approach incorporates distractor documents into the training samples. This results in k contexts for each question, comprising both true supporting documents and distractor documents.

Training the answer generation model on a dataset containing both relevant and distracting contexts allows the model to learn to distinguish between essential information and noise during inference. This enhances the model's resilience to diverse input scenarios and mitigates overfitting to specific document patterns in the training set. Additionally, this training approach aligns with the retrieval strategy, creating a harmonized training-validation environment and contributing to the model's overall robustness and adaptability.

During training, Negative Log Likelihood (NLL) loss is used to optimize the model parameters for answer generation. In this sequence-to-sequence model setting, the NLL loss can be expressed as follows: Let X be the input sequence and Y be the target sequence. The model is trained to generate the target sequence Y given the input X. The NLL loss is then calculated as the negative log likelihood of the target sequence under the model's predicted probability distribution. Assuming a token-level probability distribution $P\left(Y_i \mid X, Y_{<i}\right)$ for each token Y_i in the target sequence, the NLL loss for a single training sample is given by (3):

$$\mathcal{L}_{gen} = -\frac{1}{T}\sum_{i=1}^{T}\log P\left(Y_i \mid X, Y_{<i}\right) \tag{3}$$

where T represents the total number of tokens in the target sequence. This loss is computed for each token in the target sequence and then averaged over all tokens to obtain the overall loss for that training sample. In this paper, the input sequence refers to the input text that contains the question and contexts, while the target sequence refers to the ground truth answer to the question.

IV. EXPERIMENTS

A. DATASET

In this research, experiments were conducted on the most commonly used publicly available MMQA dataset, which is MultimodalQA. MultimodalQA [2] is an intricate QA dataset that demands comprehensive reasoning across multiple modalities, including text, tables, and images. Specifically, this dataset encompasses 16 question types, with 13 of them necessitating cross-modal retrieval and reasoning. The dataset consists of 24K question-answer pairs for training and 2.4K question-answer pairs for validation. Due to the absence of ground truth labels in the test set, the reported experimental results are based solely on the validation set. As the provided answers in the MultimodalQA dataset are mainly short and concise phrases, the metrics used for evaluating this dataset are Exact Match (EM) and average F1 score (F_1).

B. IMPLEMENTATION DETAILS

In this work, the model adopted for context retrieval is the cross-encoder [24] pre-trained on the MS Marco Passage Ranking task [25], which is ms-marco-MiniLM-L-12-v2. The base model used is a distilled version of the BERT-Base model with 12 hidden layers and a hidden size of 384 [26]. The retrieval model is trained for 5 epochs with a batch size of 32. AdamW is employed as the optimizer with a learning rate of 2e-5, and a linear scheduler is used for learning rate decay. During inference, the model is used to predict the relevance scores for all the question-document pairs, and the top-3 documents with the highest scores will be selected as the context for answer generation.

For answer generation, the model selected is the large-scale instruction fine-tuned version of the T5 model, Flan-T5-Base [27]. The model is trained for 10 epochs with an effective batch size of 8. The optimizer used is AdamW, and the initial learning rate is set to 2e-4, with a linear learning rate decay. During inference, the maximum number of output tokens is set to 50, which means that the generated answers will have a maximum length of 50 tokens. Table 1 presents an overview of the hyperparameters used during the training and fine-tuning processes.

TABLE 1. Hyperparameters used for training and fine-tuning.

	Context Retrieval	Answer Generation
Epoch	5	10
Effective Batch Size	32	8
Learning Rate	2×10^{-5}	2×10^{-4}
Optimizer	AdamW	AdamW
Scheduler	Linear	Linear

C. BASELINES

This section describes the baseline models that achieved state-of-the-art performance on the MultimodalQA dataset.

AutoRouting [2] is a method aimed at addressing questions efficiently without necessitating cross-modal reasoning. The approach involves initially identifying the modality where the answer is likely to be found and subsequently executing the corresponding single-modality module. This is achieved through a question type classifier that identifies the modality where the answer is expected, allowing for the routing of the question and relevant context to the predicted modality-specific module. The output generated by this module is then considered the final answer to the question.

ImplicitDecomp [2] is a sophisticated 2-hop implicit decomposition baseline designed to combine information from multiple modalities. It employs a RoBERTa-large-based question type classifier to predict one of 16 question types, serving as a program guiding modality selection and logical operations. During each hop, the model is supplied with the question, question type, hop number, and context associated with the relevant modality. Without explicitly decomposing questions into sub-questions, the model automatically identifies relevant parts during each hop. It uses cross-modal reasoning in the second hop by incorporating answers from the first hop, resulting in a final answer output. In contrast, for all the single-modality question types, the model exclusively uses the first hop to retrieve the answer.

Binder [53] is a novel neural-symbolic approach designed to map task inputs directly to programs without requiring training data. It offers several key features: First, Binder binds a unified API of language model functionalities to programming languages such as SQL and Python, enhancing grammar coverage to handle diverse questions effectively. Second, it adopts an LM as both the program parser and the underlying model called by the API during execution. Lastly, Binder requires only a few in-context exemplar annotations, making it efficient and versatile for solving common-sense problems. Specifically, the implementation uses GPT-3 Codex as the LM, enabling the identification of unanswerable parts in the task input and generating API calls to prompt Codex for solutions while maintaining compatibility with the original grammar structure.

Tool-interacting divide-and-conquer (TIDC) [54] is a strategy that aims to empower the collaboration of large language models (LLMs) with auxiliary tools for tackling multimodal multi-hop (MMH) question answering tasks. This method enables LLMs to break down complex MMH questions into simpler unimodal single-hop (USH) sub-questions, which will be answered using a tool specific to its modality. Distinct tools are utilized for different modalities:

Instructor-large for TextQA, TAPAS for TableQA, and BLIP-2 for ImageQA. Additionally, a Web Search tool serves as a supplement when other tools fail to extract relevant answers. Through iterative application of this strategy, LLMs can generate accurate final answers for the original MMH question. Experimentation with different LLMs revealed that the most optimal results were obtained when employing ChatGPT as the LLM alongside this strategy.

Structured Knowledge and Unified Retrieval-Generation (SKURG) [14] is a model that comprises an Entity-centered Fusion Encoder (EF-Enc) and a Unified Retrieval-Generation Decoder (RG-Dec). The EF-Enc is designed for extracting multimodal knowledge from provided sources and seamlessly integrating it using structured knowledge generated through named entity recognition (NER) and relation extraction. This integration allows for the alignment of diverse information sources into a shared semantic space, effectively reducing modality bias during knowledge retrieval and question answering. On the other hand, the RG-Dec plays a crucial role in efficiently merging intermediate retrieval outcomes into the answer generation process. Moreover, it supports adaptive retrieval step determination, which proves invaluable in navigating multi-modal and multi-hop question answering tasks with accuracy and efficiency.

Pre-trained for Reasoning Model (PReasM) [55] leverages semi-structured tables as a valuable resource for enhancing the reasoning capabilities of Language Models (LMs). The researchers have developed a synthetic dataset called D_{syn}, by scraping tables from Wikipedia. 16 distinct Example Generators (EGs) have been employed, each of which is designed for a specific reasoning skill. Multi-task pre-training was then performed using D_{syn} to improve the reasoning skills of the model. The study utilized the T5 model as the baseline and introduced the momentum sampling strategy, where samples are taken proportionately to the improving speed of the model on a task. The best result of this model is achieved when using T5-Large as the baseline model.

MMHQA-ICL [17] is an innovative framework that employs an end-to-end method to generate answers to questions. Initially, image and table data are converted into text representations, where an advanced LLaVA-based Premium Captioning Module is utilized to generate semantically enriched image captions for the image data. The framework incorporates the DeBERTa-large model as both the question type classifier and retriever, retrieving the top-3 most relevant contexts based on the question. Subsequently, a Prompt Generator Module with Type-specific In-Context Learning (ICL) Strategy generates a prompt as input for the LLM. In the reasoning stage, the text-davinci-003 API with a temperature value of 0.4 is utilized to obtain the final answer.

Solar [16] is a groundbreaking framework designed to tackle MMQA by leveraging unified language representation. This innovative approach involves transforming input tables and images into text representations, thereby simplifying the tasks into text question answering, which is easier to manage. The tables are converted into sentences linearly based on their cells, while images are processed using BLIP to generate captions and VinVL to extract attribute features. This transformation enables Solar to effectively address problems within a language space through retrieval, ranking, and generation processes. In this framework, the BERT model serves as the backbone for retrieval and ranking tasks, while the T5 model is utilized for generation purposes.

TABLE 2. Experimental results on MultimodalQA (dev-set). The best results are in bold.

Model	Single-Modal		Multi-Modal		All	
	EM	F_1	EM	F_1	EM	F_1
AutoRouting	51.7	58.5	34.2	40.2	44.7	51.1
ImplicitDecomp	51.6	58.4	44.6	51.2	48.8	55.5
Binder	-	-	-	-	51.0	57.1
TIDC-ChatGPT	-	-	-	-	43.7	61.0
SKURG	66.1	69.7	52.5	57.2	59.8	64.0
PReasM-Large	-	-	-	-	59.0	65.5
MMHQA-ICL	-	-	-	-	54.8	65.8
Solar	69.7	74.8	55.5	65.4	59.8	66.1
PERQA	69.7	74.1	54.7	60.3	62.8	67.8
UniRaG	**71.7**	**75.9**	**62.3**	**66.0**	**67.4**	**71.3**

Progressive Evidence Refinement Question & Answering (PERQA) [15] is a novel framework that adopts a two-stage architecture for multimodal retrieval question answering. In the first stage, a stepwise progressive evidence refinement strategy is employed, consisting of an Evidence Initial Screening Module (EISM) and an Iterative Evidence Retrieval Strategy (IER). This strategy focuses on selecting crucial evidence for question answering, and it introduces a negative sample semi-supervised contrastive learning training strategy to address the issue of unused distractive samples. The second stage employs a multi-turn retrieval and question answering approach, incorporating a cross-modal attention mechanism to capture connections between evidence and questions. The question answering model integrates ViT for image encoding and LLaMA with LORA (Low-Rank Adaptation) for question and text encoding, as well as decoding the combined features.

D. EXPERIMENTAL RESULTS AND PERFORMANCE COMPARISON

The performance of the proposed UniRaG framework is compared with the baseline models. Following the existing works [2], [14], [15], [16], experimental results are reported in Table 2 based on three indicators: "Single-Modal" indicates the results for samples that only require single-modal reasoning; "Multi-Modal" refers to the results for samples that necessitate reasoning across multiple modalities; and "All" encompasses the results for all the samples across the whole dataset.

From Table 2, it is evident that the proposed UniRaG framework stands out as a top performer among all the models, demonstrating its robustness and efficacy over existing state-of-the-art approaches. Before delving deeper into the analysis, it is worth noting that all methods achieved relatively higher performance in the single-modal scenarios. This may be attributed to the inherent simplicity and focused nature of processing data from a single modality, enabling the models to leverage more contextual cues within that singular modality compared to the intricacies involved in multi-modal integration. In contrast, the experimental results clearly show that all methods achieved relatively inferior performance in the multi-modal scenarios. This can be attributed to the challenges associated with integrating and synthesizing information from multiple modalities, which often lead to increased complexity and potential information loss during the retrieval and question answering processes.

Examining the experimental results in detail, UniRaG surpasses the closest competitor by a substantial margin across both single-modal and multi-modal tasks. In single-modal scenarios, the proposed framework shines by achieving an impressive EM score of 71.7% and an F1 score of 75.9%, outperforming the leading existing model (PERQA) by an appreciable margin of 2.0% in EM and 1.8% in F_1. This significant performance lead underscores the capacity of the proposed UniRaG framework to generate precise answers to questions within a single modality. This is facilitated by the incorporation of pre-trained language models in the proposed framework, which form a solid foundation for accurate comprehension and response generation.

On the other hand, in multi-modal scenarios, the proposed UniRaG framework maintains its dominance over the existing methods by achieving an EM score of 62.3% and an F1 score of 66.0%. While this performance is considerably lower compared to single-modal scenarios, UniRaG still significantly surpassed the previous state-of-the-art model (PERQA) by 7.6% EM and 5.7% F_1. This substantial performance improvement highlights the exceptional ability of the proposed framework to seamlessly integrate and process information from diverse modalities, including text, tables, and images. Specifically, the outstanding performance can be credited to the sophisticated techniques utilized for modality unification, LLaVA-v1.5-7B [22] for image-to-text transformation and table linearization for table-to-text transformation. These techniques effectively mitigate information loss and enhance the model's ability to capture nuanced relationships across different modalities, thereby improving the overall performance in multi-modal scenarios.

In the combined setting, where both single-modal and multi-modal samples are considered simultaneously, the proposed UniRaG framework excels further, achieving an EM score of 67.4% and an F1 score of 71.3%. In contrast to the prior cutting-edge model (PERQA), UniRaG attains an appreciable performance boost of 4.6% EM and 3.5% F_1. The consistent performance superiority of the proposed framework across all evaluated scenarios further accentuates its robustness and versatility in MMQA. The success of UniRaG can be attributed to its strategic framework, which adeptly unifies multimodal data into a coherent text representation before addressing MMQA using text-based methodologies. This approach bypasses the complexities inherent in directly processing diverse modalities, opting instead for a consolidated textual format that simplifies the subsequent processes. The context retrieval process is facilitated by the state-of-the-art cross-encoder pre-trained on the MS Marco Passage Ranking task, ms-marco-MiniLM-L-12-v2 [24]. Furthermore, the large-scale instruction fine-tuned version of the T5 model, Flan-T5-Base [27], is employed for answer generation. By harnessing the power of these established pre-trained language models, the proposed UniRaG framework effectively captures the intricacies of multimodal data in a unified text space, enabling it to generate contextually accurate and precise answers to the questions. Overall, this strategic integration of multimodal data into a textual framework underscores its effectiveness and versatility in tackling MMQA, positioning it as a leading solution in the research field.

E. ABLATION STUDY

The preceding section has convincingly showcased the superiority of the proposed UniRaG framework in addressing the challenges of MMQA, thereby achieving state-of-the-art performance on the MultimodalQA dataset. Here, extensive ablation studies are conducted to delve deeper into evaluating the effectiveness of the proposed framework. These experiments aim to assess the impact of novel techniques introduced in this paper, including uniform prefix conditioning, LLaVA image captioning, contextual diversity training, and context retrieval. The results obtained are presented in Table 3, which comprehensively evaluates the contribution of each technique to the overall performance on the MultimodalQA dataset. In addition to the EM and F1 scores that are used to measure the QA performance, Retrieval Recall (Retr-Rec) is incorporated into this study to evaluate the retrieval performance of the proposed UniRaG framework.

The proposed framework employs a top-k retrieval strategy, which primarily focuses on the recall rate during context retrieval. This emphasis on recall rate holds immense significance in information retrieval tasks, where the ultimate objective is to retrieve the maximum amount of relevant information. Through prioritizing the retrieval of the three most relevant documents, UniRaG showcases exceptional retrieval recall performance across various settings, encompassing single-modal, multi-modal, and overall scenarios. In the single-modal setting, UniRaG attains an outstanding recall rate of 99.0%, indicating its ability to retrieve a significantly high proportion of relevant information. In contrast, in multi-modal scenarios, where integrating and retrieving information from multiple sources poses challenges, UniRaG maintains an exceptional recall rate of 86.1%. Although the retrieval performance in the multi-modal setting is lower compared to the single-modal setting, this is reasonable given

TABLE 3. Ablation study on MultimodalQA (dev-set). The best results are in bold.

#	Model	Single-Modal			Multi-Modal			All		
		EM	F_1	Retr-Rec	EM	F_1	Retr-Rec	EM	F_1	Retr-Rec
1	UniRaG	71.7	**75.9**	99.0	**62.3**	**66.0**	**86.1**	**67.4**	**71.3**	**91.4**
2	- w/o uniform prefix conditioning	**71.9**	75.9	99.0	60.0	64.0	86.1	66.4	70.5	91.4
3	- w/o LLaVA image captioning	69.0	73.4	**99.2**	53.3	56.9	80.5	61.8	65.8	88.2
4	- w/o contextual diversity training	60.1	64.0	99.0	52.5	55.9	86.1	56.6	60.3	91.4
5	- w/o context retrieval	24.6	26.6	14.6	16.2	18.9	13.1	20.7	23.1	13.7

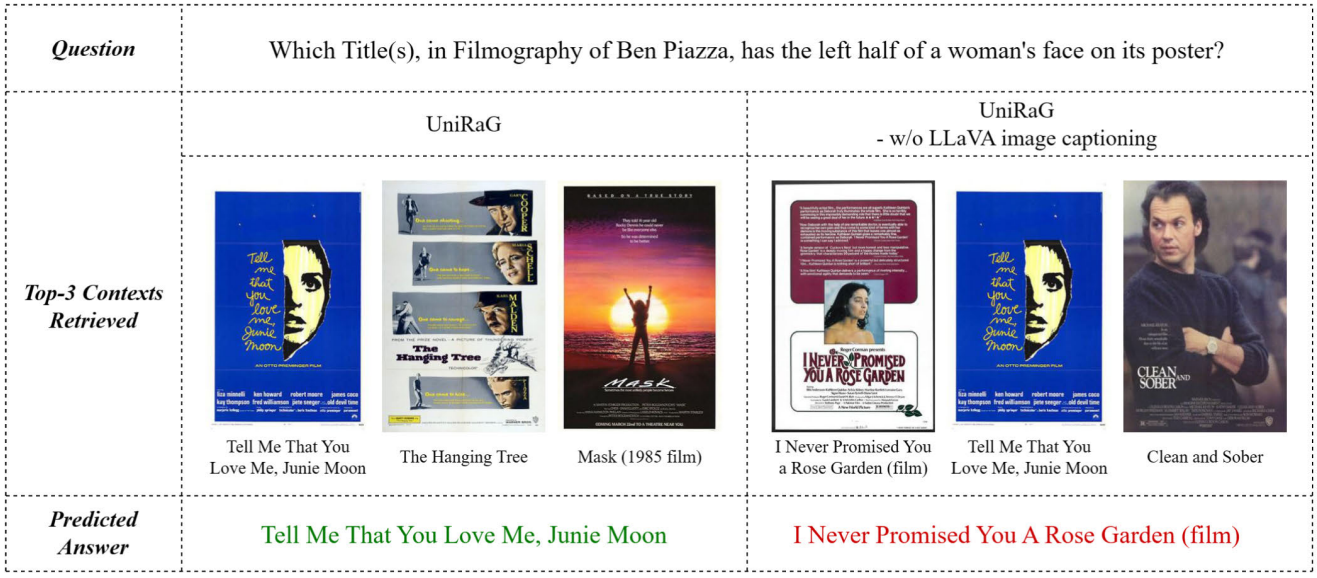

FIGURE 3. Example with and without LLaVA image captioning.

the inherent complexity of multimodal information retrieval. Furthermore, when considering the overall performance across all modalities, the proposed UniRaG framework consistently demonstrates exceptional recall rates of 91.4%, thereby showcasing its robustness and effectiveness in comprehensively retrieving relevant data across diverse scenarios.

The second row presents the results obtained when the answer generation model is trained without employing the uniform prefix conditioning strategy. By omitting this fine-tuning approach, the Flan-T5-Base model is trained by simply concatenating the question and contexts as input text for answer generation. Although this strategy does not have much impact in the single-model scenario, a noticeable decrease in performance has been observed in the multi-modal scenario (−2.3% EM, −2.0% F_1), leading to suboptimal performance in the overall MMQA task. From another perspective, the third row displays the results achieved when the advanced LLaVA image captioning is not utilized during the image-to-text transformation. Instead, the widely used image captioning model, BLIP, is employed to convert images into textual representations. Consequently, there has been a significant drop in overall performance across all metrics on MultimodalQA dataset (−5.6% EM, −5.5% F_1, −3.2% Retr-Rec). This result underscores the effectiveness of LLaVA in mitigating information loss during the image-to-text transformation, thereby facilitating the generation of more contextually accurate and comprehensive image descriptions that are crucial for context retrieval and question answering tasks.

Following that, the fourth row showcases the results obtained when the answer generation model does not undergo context diversity training. In this case, the model is trained exclusively using the relevant documents, where the distractor documents are not included in the training samples. In this case, the model may encounter challenges in extracting pertinent information from extraneous sources, resulting in unsatisfactory QA performance. This was evidenced in the ablation results shown in Table 3, where the QA performance of the model dramatically declined across all scenarios. Specifically, the EM score and F1 score dropped by 11.6% and 11.9% in single-modal scenarios and decreased by 9.8% and 10.1% in multi-modal scenarios, resulting in the overall degradation of 10.8% EM and 11.0% F_1. Lastly, the fifth row presents the results with the removal of the context retrieval stage, in which random documents are selected as the contexts for answer generation. As a result, the model struggled to generate precise answers to the question, attaining significantly poor performance on the

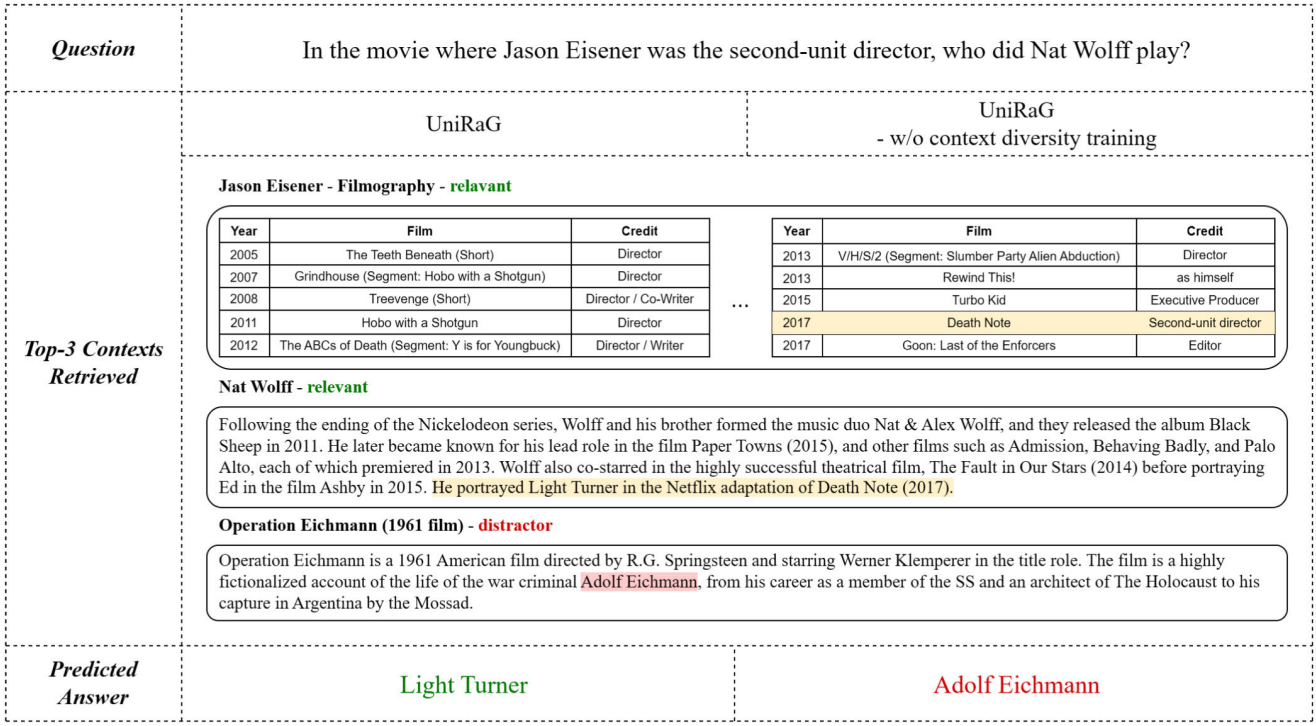

FIGURE 4. Example with and without context diversity training.

MultiModalQA dataset. Hence, this finding highlights the critical importance of effective context retrieval in addressing the MMQA challenges.

F. CASE STUDY

In this section, a deeper evaluation is conducted to assess the effectiveness of the novel techniques introduced in the proposed UniRaG framework through the presentation of several illustrative examples. Figure 3 shows a comparison between the results obtained with and without the implementation of LLaVA image captioning during the image-to-text transformation process. Although the true context (''Tell Me That You Love Me, Junie Moon'') can be retrieved in both cases, the model fails to produce the correct final answer when LLaVA image captioning is not utilized. The omission of LLaVA image captioning leads to a loss of critical information during the transition from image to text, resulting in the model's inability to accurately generate the correct answer.

Figure 4 provides a comparative analysis of the outcomes achieved with and without context diversity training within the proposed UniRaG framework. Given the identical contexts, the model devoid of context diversity training struggles to distinguish between pertinent information and distractor documents, thereby generating incorrect answers to the questions posed. In stark contrast, the incorporation of context diversity training empowers the UniRaG to effectively identify the relevant contexts and decipher intricate contextual relationships between them, thereby facilitating the generation of precise and correct final answers.

V. CONCLUSION

In this paper, a comprehensive three-stage framework, UniRaG, is specifically proposed to address MMQA, involving unified knowledge representation, context retrieval, and answer generation. By harnessing the capabilities of PLMs, UniRaG has demonstrated remarkable performance in MMQA, excelling in both retrieval and question answering. The multimodal knowledge is seamlessly integrated into a unified textual representation at the initial stage, where LLaVA image captioning is utilized to generate rich and detailed descriptions for the images and the table linearization technique is used to convert tabular data into textual representations. Subsequently, a cross-encoder pre-trained on the MS Marco Passage Ranking task, ms-marco-MiniLM-L-12-v2, is further fine-tuned on sequence classification to predict the relevance scores for question-document pairs, thereby selecting the top-k documents with the highest scores as the contexts for answer generation. Finally, the answer generation stage is supported by leveraging the state-of-the-art PLM, Flan-T5-Base, which has shown preferable performance across various NLP tasks. In this stage, uniform prefix conditioning and contextual diversity training are introduced to further improve the robustness of the model, thereby facilitating enhanced question answering performance. Through extensive experimentation and validation, the superior performance of the proposed UniRaG

framework on the MultimodalQA dataset is demonstrated, solidifying its effectiveness and reliability in tackling the challenges of MMQA. In a nutshell, UniRaG represents a significant advancement in multimodal comprehension and question answering research. Moving forward, continued exploration and refinement in MMQA hold the promise of unlocking further advancements and insights, leading to the development of more sophisticated and comprehensive multimodal AI systems.

REFERENCES

[1] D. Hannan, A. Jain, and M. Bansal, "ManyModalQA: Modality disambiguation and QA over diverse inputs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 7879–7886.

[2] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant, "MultiModalQA: Complex question answering over text, tables and images," 2021, *arXiv:2104.06039*.

[3] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1533–1544.

[4] Y. Yang, W.-T. Yih, and C. Meek, "WikiQA: A challenge dataset for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2013–2018.

[5] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," 2016, *arXiv:1606.05250*.

[6] P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," 2015, *arXiv:1508.00305*.

[7] S. Vakulenko and V. Savenkov, "Tableqa: Question answering on tabular data," *CoRR*, vol. abs/1705.06504, 2017.

[8] L. Nan, C. Hsieh, Z. Mao, X. V. Lin, N. Verma, R. Zhang, W. Kryscinski, H. Schoelkopf, R. Kong, X. Tang, M. Mutuma, B. Rosand, I. Trindade, R. Bandaru, J. Cunningham, C. Xiong, D. Radev, and D. Radev, "FeTaQA: Free-form table question answering," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 35–49, Jan. 2022.

[9] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[11] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6693–6702.

[12] Y. Li, W. Li, and L. Nie, "MMCoQA: Conversational question answering over text, tables, and images," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 4220–4231.

[13] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text," 2022, *arXiv:2210.02928*.

[14] Q. Yang, Q. Chen, W. Wang, B. Hu, and M. Zhang, "Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5223–5234.

[15] S. Yang, A. Wu, X. Wu, L. Xiao, T. Ma, C. Jin, and L. He, "Progressive evidence refinement for open-domain multimodal retrieval question answering," 2023, *arXiv:2310.09696*.

[16] B. Yu, C. Fu, H. Yu, F. Huang, and Y. Li, "Unified language representation for question answering over text, tables, and images," 2023, *arXiv:2306.16762*.

[17] W. Liu, F. Lei, T. Luo, J. Lei, S. He, J. Zhao, and K. Liu, "MMHQA-ICL: Multimodal in-context learning for hybrid question answering over text, tables and images," 2023, *arXiv:2309.04790*.

[18] H. Luo, Y. Shen, and Y. Deng, "Unifying text, tables, and images for multimodal question answering," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 1–13.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[22] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023, *arXiv:2310.03744*.

[23] P. Li, Y. He, D. Yashar, W. Cui, S. Ge, H. Zhang, D. R. Fainman, D. Zhang, and S. Chaudhuri, "Table-GPT: Table-tuned GPT for diverse table tasks," 2023, *arXiv:2310.09263*.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

[25] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, "MS MARCO: A human generated machine reading comprehension dataset," 2016, *arXiv:1611.09268*.

[26] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5776–5788.

[27] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, and S. Brahma, "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[28] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3190–3199.

[29] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, "KVQA: Knowledge-aware visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8876–8884.

[30] W. Chen, H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Wang, "HybridQA: A dataset of multi-hop question answering over tabular and textual data," 2020, *arXiv:2004.07347*.

[31] W. Chen, M.-W. Chang, E. Schlinger, W. Wang, and W. W. Cohen, "Open question answering over tables and text," 2020, *arXiv:2010.10439*.

[32] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua, "TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance," 2021, *arXiv:2105.07624*.

[33] Y. Chang, G. Cao, M. Narang, J. Gao, H. Suzuki, and Y. Bisk, "WebQA: Multihop and multimodal QA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16474–16483.

[34] R. G. Reddy, X. Rui, M. Li, X. Lin, H. Wen, J. Cho, L. Huang, M. Bansal, A. Sil, and S.-F. Chang, "Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 11200–11208.

[35] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[36] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23318–23340.

[37] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[38] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26183–26197.

[39] L. H. Suadaa, H. Kamigaito, K. Funakoshi, M. Okumura, and H. Takamura, "Towards table-to-text generation with numerical reasoning," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process*, 2021, pp. 1451–1465.

[40] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instructional Conf. Mach. Learn.*, 2003, vol. 242, no. 1, pp. 29–48.

[41] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[42] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," 2020, *arXiv:2004.04906*.
[43] W.-C. Chang, F. X. Yu, Y.-W. Chang, Y. Yang, and S. Kumar, "Pre-training tasks for embedding-based large-scale retrieval," 2020, *arXiv:2002.03932*.
[44] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2020, pp. 39–48.
[45] R. Nogueira and K. Cho, "Passage re-ranking with BERT," 2019, *arXiv:1901.04085*.
[46] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "CEDR: Contextualized embeddings for document ranking," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 1101–1104.
[47] R. Nogueira, Z. Jiang, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," 2020, *arXiv:2003.06713*.
[48] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[49] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," 2021, *arXiv:2111.09543*.
[50] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–25.
[51] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, and E. Xing, "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–29.
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[53] Z. Cheng, T. Xie, P. Shi, C. Li, R. Nadkarni, Y. Hu, C. Xiong, D. Radev, M. Ostendorf, L. Zettlemoyer, N. A. Smith, and T. Yu, "Binding language models in symbolic languages," 2022, *arXiv:2210.02875*.
[54] H. Rajabzadeh, S. Wang, H. Ju Kwon, and B. Liu, "Multimodal multi-hop question answering through a conversation between tools and efficiently finetuned large language models," 2023, *arXiv:2309.08922*.
[55] O. Yoran, A. Talmor, and J. Berant, "Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills," 2021, *arXiv:2107.07261*.

QI ZHI LIM received the bachelor's degree (Hons.) in computer science (artificial intelligence) from Multimedia University, Malaysia, in 2023, where he is currently pursuing the Ph.D. degree in IT, with a focus on multimodal question answering. His research interests include multimodal data pre-processing, feature extraction, information retrieval, and question-answering.

CHIN POO LEE (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in information technology in the area of abnormal behavior detection and gait recognition. She is currently an Associate Professor with the Faculty of Information Science and Technology, Multimedia University, Malaysia. Her research interests include natural language processing, gait recognition, action recognition, computer vision, and deep learning.

KIAN MING LIM (Senior Member, IEEE) received the B.I.T. degree (Hons.) in information systems engineering and the Master of Engineering Science (M.Eng.Sc.) and Ph.D. degrees in IT from Multimedia University. He is currently an Associate Professor with the Faculty of Information Science and Technology, Multimedia University. His research interests include machine learning, deep learning, computer vision, and pattern recognition.

AHMAD KAMSANI SAMINGAN received the Ph.D. degree from the University of Southampton, U.K., in 2004. He has been appointed by the Ministry of Energy, Science, Technology and Innovation (MOSTI) as an Expert Panelist for evaluating proposals for the ministry's research and development grants and monitoring undergoing research and development projects. He is currently a Solution Architect with Telekom Research and Development, specializing in artificial intelligence and autonomous solutions. He is also an Adjunct Professor with the Faculty of Electric and Electronic Engineering, Universiti Tun Hussin Onn Malaysia (UTHM). He has filed 11 patents and published 25 research journals and international conference papers. He has more than 20 years of experience in research and development. His research interests include wireless systems (5G and beyond), antenna and channel propagation, autonomous systems (predictive, prescriptive, and preemptive), data analysis, natural language processing (NLP), automated speech recognition (ASR), machine learning, and artificial intelligence. He has been involved in many research and development projects in his career, which he served as the project leader for most of them. Some of his project's outputs have received national and international recognition, such as the Best of Communication Merit Award at APICTA Malaysia and APICTA International, in 2017, the WITSA Merit Award, in 2018, and the Winner of the Malaysian Technology Excellence Awards, in 2022, for connectivity—telecommunications category.