MAR: Matching-Augmented Reasoning for Enhancing Visual-based Entity Question Answering

Zhengxuan Zhang[1], **Yin Wu**[1], **Yuyu Luo**[1,2], **Nan Tang**[1,2*]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
{zzhang393, ywu450}@connect.hkust-gz.edu.cn {yuyuluo, nantang}@hkust-gz.edu.cn

Abstract

A multimodal large language model (**MLLM**) may struggle with answering *visual-based (personal) entity questions* (**VEQA**), such as "who is A?" or "who is A that B is talking to?" for various reasons, *e.g.,* the absence of the name of A in the caption or the inability of **MLLMs** to recognize A, particularly for less common entities. Furthermore, even if the MLLM can identify A, it may refrain from answering due to privacy concerns. In this paper, we introduce a novel method called Matching-Augmented Reasoning (**MAR**) to enhance **VEQA**. Given a collection of visual objects with captions, **MAR** preprocesses each object individually, identifying faces, names, and their alignments within the object. It encodes the information and stores their vector representations in the database. When handling **VEQA**, **MAR** retrieves matching faces and names and organizes these entities into a matching graph. **MAR** then derives the answer to the query by reasoning over this matching graph. Extensive experiments show that **MAR** significantly improves **VEQA** compared with the state-of-the-art methods using **MLLMs**.[1]

1 Introduction

Multimodal language models (**MLLMs**) (Cui et al., 2024) like GPT-4V (Zhang et al., 2023a) and LLaVA (Liu et al., 2023) have significantly improved visual question answering (VQA) by integrating text and images. However, they still face challenges in visual-based entity question answering (**VEQA**), a crucial subset of VQA that focuses on extracting information about specific entities (Qiu et al., 2024; Chen et al., 2023a).

MLLMs for VEQA: Advantages and Limitations. In **VEQA** tasks, **MLLMs** excel at integrating visual cues and textual information for effective reasoning and answer generation (Li et al., 2023b; Liu et al., 2024b). For instance, as depicted in Figure 1(a), GPT-4V, when tasked with answering question Q_1 regarding the face in region R_1, leverages the associated caption T_1 of image V_1 to precisely identify the person within the red box as "Wang Yi".

However, **MLLMs** often struggle to recognize all details in images, particularly for less common entities (Li et al., 2023b; Sun et al., 2024; Yang et al., 2024; Wu et al., 2024). For instance, in Figure 1(b), GPT-4V fails to answer question Q_2 about the person in the red rectangle R_2 due to the lack of information in the image caption T_2 and its limited knowledge base. Furthermore, even when an **MLLM** identifies an entity, it may withhold an answer due to privacy regulations.

Figure 1: **Data** (V : image, T : text) pair; **Query** (R : entity selection, Q : question) pair. (a) The advantages of **MLLMs**; (b) The limitations of **MLLMs**, and (c) Our proposal **MAR**.

*Nan Tang is the corresponding author
[1]Our dataset and method are publicly available at https://github.com/HKUSTDial/MAR.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1520–1530
November 12-16, 2024

Despite rapid advancements of **MLLMs**, accurately identifying all personal entities in images and adhering to privacy regulations make answering **VEQA** questions solely using **MLLMs** a significant challenge (Chen et al., 2024; Li et al., 2023a, 2024b; Yu et al., 2023; Qin et al., 2022).

Matching-Augmented Reasoning (MAR). Given a collection of visual objects with captions, sourced from public or enterprise datasets without privacy concerns, **MAR** identifies the faces of entities within visual objects and the names of entities within captions by tools like CLIP (Radford et al., 2021) and Deepface (Taigman et al., 2014). These entities are encoded with respective visual and text encoders, and the resulting embeddings are stored in vector databases *e.g.,* Meta Faiss (Douze et al., 2024). When a **VEQA** query is posed, **MAR** retrieves "similar" faces and names from the database and performs reasoning over these matched pieces of information to generate an accurate response.

Existing work on **VEQA** (Chen et al., 2023a; Hu et al., 2023; Qiu et al., 2024) mainly focuses on general entities such as animals buildings, and vehicles. However, there is a lack of work targeting personal entities. As illustrated in Figure 1(c), if we can match the face in image V_2 with the face in image V_1, and if we know that the face in V_1 is "Yi Wang", we can answer Q_2.

Contributions. We notable contributions are summarized as follows.

- We study **VEQA**, an important and commonly used subset of VQA, but it is not fully explored.
- We propose *matching graphs* that can capture the relationships of the same entities over multiple captioned visual objects. Based on a matching graph, we proposed matching-augmenting reasoning (**MAR**), to effectively answer a **VEQA**.
- Given the lack of **VEQA** dataset focusing on the personal entity, we construct a new benchmark **NewsPersonQA** including $235k$ images and $6k$ QA pairs.
- We conduct extensive experiments to show that **MAR** $>$ **MLLMs** + **RAG** $>$ **MLLMs**, where **RAG** is to feed the retrieved matching graph to **MLLMs**.

The structure of our paper is organized as follows: Section 1 introduces the limitations of using **MLLMs** to answer visual questions and proposes the **VEQA** task. Section 2 reviews related work on the **VEQA** task. In Section 3, we provide a detailed description of **VEQA**. Section 4 is dedicated to presenting our approach, **MAR**, for addressing this task. Section 5 presents the benchmark **NewsPersonQA** we proposed, and Section 6 describes extensive experiments conducted to validate our approach. Finally, Section 7 summarizes the findings and contributions of our paper.

2 Related Work

We categorize related work as follows.

2.1 Visual Question Answering (VQA)

VQA aims at reasoning over visual and textual content and cues to generate answers (Lu et al., 2021; Stengel-Eskin et al., 2022; Agrawal et al., 2023). It primarily utilizes approaches such as Fusion-based (Zhang et al., 2019), Multimodal Learning (Ilievski and Feng, 2017), Memory Networks (Su et al., 2018), Visual Attention (Mahesh et al., 2023), etc., to discover and integrate information from text and images.

2.2 Multimodal Large Language Models (MLLMs) for VQA

MLLMs, such as GPT-4V (Zhang et al., 2023a) and LLaVa (Liu et al., 2023), have played a pivotal role in advancing VQA. By seamlessly integrating textual and visual information, these models have demonstrated a remarkable ability to understand and respond to complex queries about images.

2.3 Retrieval-Augmented Generation (RAG) for VQA

In many cases, the cues within images and text are insufficient for reasoning and answering. Retrieval-augmented generation (RAG) (Lewis et al., 2021; Chen et al., 2023b; Li et al., 2024a; Liu et al., 2024a) has been studied for VQA, especially with Knowledge-Based VQA approaches that incorporate external knowledge to provide additional cues for answers (Khademi et al., 2023; Shah et al., 2019).

2.4 Visual-based Entity Question Answering (VEQA)

Recent advancements in VQA (Qiu et al., 2024; Chen et al., 2023a; Hu et al., 2023) have focused

on entity-based questions involving general entities like buildings and animals, while personal entities remain unexplored. **MLLMs** struggle with questions about human entities due to limited knowledge and privacy issues (Section 6). Although RAG (Tang et al., 2024) can enhance **MLLMs** for **VEQA** tasks, challenges persist in reasoning with multiple interconnected visual objects.

2.5 Data Matching

This involves identifying, comparing, and merging records from multiple datasets to determine duplicate entities (Tu et al., 2023; Ebraheem et al., 2018; Xie et al., 2024). With increasing data multimodality, matching has expanded from string matching (Text-Text) and entity matching (Tuple-Tuple) to include Image-Text (Li et al., 2019; Mai et al., 2023; Zhang et al., 2023b) and Image-Image (Zhu et al., 2018) matching. Matching aggregates clues, enhances model reasoning, and offers strong interpretability (Zheng et al., 2022).

3 Visual-based Entity Question Answering (VEQA)

Captioned Visual Objects. We consider a *captioned visual object* O as a pair $O : (V, T)$ where V is an image, and T is an optional text description relative to the image V.

Figure 1(a) and Figure 1(b) provide two sample captioned visual objects, (V_1, T_1) and (V_2, T_2), respectively.

Let $\mathbf{O} = \{O_1, O_2, \ldots, O_n\}$ be a group of captioned visual objects, sourced from public or enterprise datasets without privacy concerns. Note that, such a group is common in practice, *e.g.,* a collection of news articles.

VEQA. Users can pose a Visual-based Entity Question Answering (**VEQA**) queries related to person entities on either a single captioned visual object (Single-**VEQA**) or a group of such objects (Group-**VEQA**).

Single-VEQA. Given a captioned visual object $O : (V, T)$, this type of queries allows the user to provide a rectangle selection of the image and ask the question like "who is he/she".

More formally, a Single-**VEQA** $\mathbf{Q}_s$ is a pair (R, Q), where R is a rectangle selection over image V and Q is a natural language question.

Group-VEQA. Given a group of captioned visual objects $\mathbf{O}$, we support two types of queries $\mathbf{Q}_g$: (1) a simple natural language query Q, such as "how many news contain Donald Trump"; and (2) a natural language query with a selected face, *i.e.,* a pair (R, Q), such as "in which news the selected person appears".

We will simply use **Q** to represent either a Single-**VEQA** or a Group-**VEQA** query.

4 Algorithms for VEQA

Next, we will first discuss using **MLLMs** for **VEQA** in Section 4.1, and then discuss coarse-grained RAG in Section 4.2. We then propose a new concept "matching graphs" that provides fine-grained information among retrieved objects in Section 4.3, based on which we describe fine-grained RAG in Section 4.4 and matching-augmented reasoning (**MAR**) in Section 4.5.

4.1 MLLMS for VEQA

Given a **VEQA** query **Q**, a crude solution is to directly prompt **Q** to a **MLLM** as:

Q $\rightarrow$ **MLLM** $\rightarrow$ answer

Figure 2(a) depicts this solution.

4.2 Coarse-Grained RAG for VEQA

Alternatively, we can retrieve top-k captioned visual objects and feed them to **MLLMs** as:

(**Q**, top-k objects) $\rightarrow$ **MLLM** $\rightarrow$ answer

Figure 2(b) illustrates this approach, which we refer to as *coarse-grained RAG*. This method is characterized by its transmission of entire retrieved objects to the **MLLMs**. Unfortunately, current **MLLMs** perform poorly in reasoning with multiple interconnected retrieved visual objects.

4.3 Matching Graphs

To improve the performance of RAG models, it's beneficial to focus on fine-grained information rather than entire objects. By identifying specific entities (*e.g.,* faces, names) and their connections within each object, we can provide a more meaningful context for reasoning.

Matching Graphs. A matching graph $G(N, E)$ contains a set N of nodes and a set E of undirected edges. Each node $n \in N$ has two labels **face**(n) and **name**(n), where **face**(n) is a face image, and **name**(n) is a set of possible names.

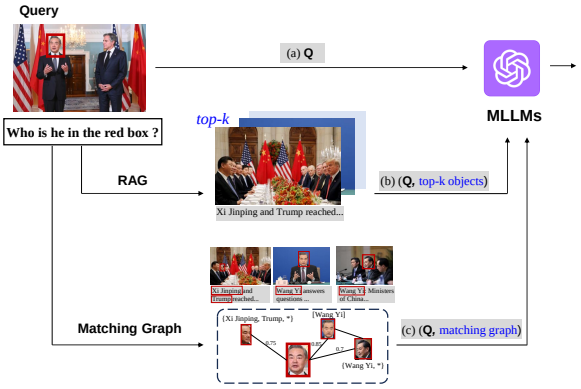

Figure 2: Different algorithms for **VEQA**. (a) **MLLMs**. (b) Coarse-grained RAG. (c) Fine-grained RAG.

If we are certain about a person's name, we will use a square bracket *e.g.*, $\texttt{name}(n) = [\text{Yi Wang}]$ for the selected face in Figure 1(a); if we are not sure about a person's name, we will use a curly bracket to indicate possible names *e.g.*, $\texttt{name}(n) = \{\text{Xi Jinping, Trump, *}\}$ for the selected face in Figure 1(b), where $*$ is a wildcard meaning that n's name could be something other than Xi Jinping and Trump.

Each undirected edge $e(n_i, n_j) \in E$ indicates that the two faces corresponding to n_i (*i.e.*, $\texttt{face}(n_i)$) and n_j (*i.e.*, $\texttt{face}(n_j)$) are likely to be the same person. Each edge has a weight $\texttt{weight}(e) \in [0, 1]$, indicating the *similarity* of the two faces.

Matching Graph Construction. It consists of two steps: offline index construction (for all data objects) and online matching graph construction (for each query).

Offline Index Construction. We first preprocess each captioned visual object $O(V, T)$ as follows.

- **Face identification.** We use Meta DeepFace (Taigman et al., 2014) to extract face entities as $(f_1, f_2, \ldots, f_k)$ from image V.
- **Name identification.** We use spaCy (Honnibal et al., 2020) to extract name entities as $(x_1, x_2, \ldots, x_m)$ from text T.

After pre-processing, we have constructed all possible nodes for all possible matching graphs. We then use pre-trained CLIP (Radford et al., 2021) to convert each identified face and each identified person names into its vector representation, and store them in two separate vector database: **faceDB** and **nameDB**.

Iterative Online Matching Graph Construction. Given a **VEQA** query, we construct a matching graph as follows.

[**Step 1:** Initialization.] The user starts with a *seed node* (for Single-**VEQA**) or a group of *seed nodes* for (Group-**VEQA**). Each seed node contains a face and its candidate names that could be empty.

[**Step 2:** Graph Expansion.] For each node in the graph, we search either similar faces from **faceDB** with vector similarity above a given threshold σ_f, or similar names from **nameDB** with vector similarity above a given threshold σ_n. For each added node, the edge weight is set as face similarity.

[**Step 3:** Iterative Search and Termination.] When there are new nodes added in Step 2, we will loop Step 2. The process terminates when either there is no new nodes can be added or we have done k iterations. From our empirical findings, we set $k = 2$, which is enough to retrieve useful nodes (*e.g.*, 10 nodes) and edges for reasoning.

4.4 Fine-Grained RAG for VEQA

Given the fine-graph matching graph relative to a query **Q**, we prompt it to **MLLMs** as:

(**Q**, matching graph) → **MLLM** → answer

Figure 2(c) shows this approach, which we refer to as *fine-grained RAG*. It works as follows.

[**Step 1:** Image Stitching.] Most **MLLMs** (*e.g.*, LLaVA) only support only single-image input, thus we simply combine multiple retrieved visual objects into one visual object **V**.

[**Step 2:** Image Annotation.] We annotate each node n_i in the matching graphs on **V** in a red box, resulting in an annotated image $\mathbf{V}'$.

[**Step 3:** Matching Graph Serialization.] Each node n_i and edge $e(n_i, n_j)$ will be serialized as:

$$\texttt{ser}(n_i) = \texttt{face}(n_i), \texttt{name}(n_i)$$
$$\texttt{ser}(e) = n_i, n_j, \texttt{weight}(e)$$

Serializing a matching graph $g(N, E)$ is to serialize all nodes and edges as:

$$\texttt{ser}(g) = \texttt{ser}(N), \texttt{ser}(E)$$

We then prompt **Q**, $\mathbf{V}'$, and $\texttt{ser}(g)$ to **MLLMs**. In order to enable it to consider information from its own model simultaneously, we also designed an Original knowledge-aware Prompt (OP): "*Please tell me* [**Q**]. *If you are unsure, read the following.*"

4.5 MAR for VEQA

MAR for Single-VEQA. This type of queries asks the name of a single entity. Given a matching graph $g(N, E)$ where $n^* \in N$ is the seed node, our method works as follows.

[**Step 1:** Remove Uncertain Nodes.] For each node $n_i \in N \setminus \{n^*\}$, if its name is uncertain, we remove n_i and its associated edges, which will resulted in a modified graph $g(N', E')$.

[**Step 2:** Name Aggregation for n^*.] We count all distinct names in the modified matching graph g', each associated with a weight as $\sum_{e(n_i,n^*)\in E'} \texttt{weight}(e)$.

[**Step 3:** Name Identification for n^*.] We pick the name with the highest weight, as the answer to the Single-**VEQA** query.

MAR for Group-VEQA. This type of queries ask for aggregated information of nodes whose names are queried in the query, *e.g.,* "which image/how many images have person A". Given a matching graph $g(N, E)$, it works as follows.

[**Step 1:** Name Identification for Each Node.] It first identifies the name of each node, as discussed above.

[**Step 2:** Answer Aggregation.] It aggregates the information of each node to answer the given Group-**VEQA**.

5 A New NewsPersonQA Benchmark

The problem of **VEQA** needs to address complex interactions between multiple visual and textual data. Despite its growing importance, existing benchmarks fall short in adequately representing the diverse challenges posed by **VEQA** tasks. Particularly in the domain of News QA, where the accurate identification and understanding of both common and uncommon persons are crucial, current datasets (*e.g.,* GoodNews (Biten et al., 2019) and NewsQA (Trischler et al., 2016)) do not provide the necessary depth and breadth. To bridge this gap, based on GoodNews (Biten et al., 2019), we are constructing a new benchmark, namely **NewsPersonQA**, that encompasses a wide range of scenarios, including both well-known and obscure individuals.

Table 1: Statistics of **NewsPersonQA**

Category	Count
Total Images	235,912
Totally Extracted Faces	336,075
Totally Extracted Names	379,313
Single-**VEQA** Queries	4,937
Group-**VEQA** Queries	1,004
Total Queries	**5,941**

5.1 The construction of the dataset

The construction of the dataset entails the generation of QA pairs from the raw data in GoodNews, which consists of images and captions. This process involves two main steps: data preprocessing and QA pair construction.

Data Preprocessing: Raw data undergoes preprocessing, which includes structuring news data, extracting faces from images, annotating original images, and recognizing named entities in captions. The processed data is then randomly distributed into groups. Each group contains thousands of images and is categorized into Single-**VEQA** (100 groups) and Group-**VEQA** (10 groups) queries.

Single-VEQA Question Generation: We begin by counting the frequency of each person's name within each group. To ensure the availability of clues for answering, we select names that appear at least three times in captions. We then mask these names in the captions to generate QA pairs. For example: **Question:** "Who is the person labeled 'face n' in the red box?" **Answer:** "name". In total, approximately 5,000 queries of this type are generated, about 50 per group.

Group-VEQA Question Generation: Similarly, we count the occurrences of names within each group and store the image names as a set, denoted as S. To prevent exceeding the maximum token limit of **MLLMs** in the answers and to facilitate clearer visualization of experimental results, we limit each person's name to a maximum of 5 appearances within the same group. We then randomly mask part of the captions corresponding to the images in the set to increase the difficulty and encourage **MLLMs** to generate correct answers through retrieved content. The format of QA pairs is **Question:** "Which photos are of the person named 'name'?" **Answer:** S. The number

of queries of this type is approximately 1,000.

Table 1 shows the statistics of **NewsPersonQA**.

5.2 Comparison between Existent VEQA Datasets and NewsPersonQA

In recent years, numerous VEQA datasets and methods have been developed, including OVEN (Hu et al., 2023), INFOSEEK (Chen et al., 2023a), and SnapNTell (Qiu et al., 2024). Our discussion primarily focuses on these works.

Different Types of Entities: These works mainly focus on general entities, such as buildings, animals, and vehicles, and do not address personal entities. Person entities are an important type of entity. However, due to privacy policies and other reasons, some **MLLMs** (such as GPT-4V, Claude, etc.) cannot directly answer questions related to person entities, thus leaving a gap that needs to be filled.

Different Dataset Division Structures: Previous works primarily aim to enable models to learn relevant knowledge through training and then perform testing. Therefore, their datasets are divided into training, validation, and test sets. Unlike them, our work aims to assist VEQA by allowing the model to find relevant clues in the database through a zero-shot approach. Thus, our dataset is divided based on the database, and the model is tasked with finding clues within a specific database.

6 Experiment

Methods. For answering **VEQA** queries, we selected two well-known and highly capable **MLLMs**, as well as human evaluation,to serve as baselines.

- **LLaVA:** This model utilizes CLIP-ViT-L-336px with an MLP projection. We refer to the 1.5 version with 7 billion parameters as LLaVA-7b and the version with 13 billion parameters as LLaVA-13b.
- **GPT-4V:** Recognized as OpenAI's most powerful general-purpose MLLM to date, GPT-4V boasts 1.37 trillion parameters.
- **Human:** This represents the human-annotated results, showcasing the level of cognitive ability and performance that humans can achieve on this task.

Table 2: Result for Singe-**VEQA** Queries. (Note: GPT-4V could not answer these queries directly due to policy constraints. Values within parentheses are those GPT-4V still refuses to answer.)

Models	Acc (%)	Acc^{hit} (%)
Human	3.36	5.19
Human + FRAG	**47.01**	**98.31**
LLaVA-7b	22.26	27.53
LLaVA-7b + FRAG	31.19	62.81
LLaVA-13b	27.93	32.86
LLaVA-13b + FRAG	31.13	62.34
GPT-4V	-	-
GPT-4V + FRAG	34.84 (4.2)	68.31 (2.6)
MAR	**39.09**	**79.65**

Table 3: Result for Group-**VEQA** Queries.

Models	Recall
LLaVA-7b + FRAG	22.06%
LLaVA-13b + FRAG	40.05%
GPT-4V + FRAG	65.04%
MAR	**70.85%**

+ **FRAG:** **MLLMs** struggle with reasoning over coarse-grained RAG that consists of multiple captioned visual objects. Therefore, we provide only fine-grained RAG (FRAG), *i.e.,* matching graph, to the above-mentioned models and human evaluators.

Implementation. The experiments were conducted in a zero-shot setting using RTX 4090 GPUs. For GPT-4V, we used the interface of the GPT-4-vision-preview model. It's worth noting that GPT-4V often refrains from answering person identify questions without additional clues due to policy reasons. However, with the incorporation of matching graph techniques, it can leverage weak signals and combine them with its own knowledge base. In the case of Group-**VEQA** queries, a maximum of 10 cases are recalled and then filtered for subsequent processing.

Metrics. For Single-**VEQA** queries, we use accuracy (**Acc**) as an evaluation metric. Furthermore, we assess the accuracy only for instances where relevant clues are successfully retrieved (*e.g.,* the case of Figure 1(c)), which is denoted as $\textbf{Acc}^{\textbf{hit}}$. For Group-**VEQA** queries, we employ recall (**Recall**) as the metric.

Table 4: Study on Successfully Recalled Data.

Models	$\text{Acc}^{\texttt{hit}}$ (%)
LLaVA-7b	
w/o FRAG ✘ → with FRAG ✓	42.86
w/o FRAG ✓ → with FRAG ✘	7.32
LLaVA-13b	
w/o FRAG ✘ → with FRAG ✓	39.18
w/o FRAG ✓ → with FRAG ✘	9.44

Table 5: Ablation Study: Original-knowledge-aware Prompt (OP)

Models	Acc
LLaVA-7b with matching	31.19%
w/o OP	25.14%
LLaVA-13b with matching	31.13%
w/o OP	29.41%
GPT-4V with matching	39.09%
w/o OP	34.58%

6.1 Single-VEQA Queries

The main results from the Single-**VEQA** queries are summarized in Table 2, which leads to the following insights:

1. Model Parameter Size: LLaVA-13b demonstrates higher accuracy (27.93%) compared to LLaVA-7b (22.26%), suggesting that a model's recognition ability is positively correlated with its parameter size, which to some extent reflects its knowledge base.

2. Impact of Matching Graph: Incorporating a matching graph leads to an 8.9% improvement in accuracy for LLaVA-7b and a 3.2% improvement for LLaVA-13b. GPT-4V, with matching, achieves a character recognition accuracy of 34.83%.

3. Comparative Improvement: The enhancement from matching is more pronounced for LLaVA-7b than for LLaVA-13b, indicating that while matching can compensate for differences in parameters, a model's inherent capabilities still set an upper limit on its performance.

To further understand the impact of matching on the **models' reasoning abilities**, we analyzed examples of successfully recalled clues:

i. Human Performance: Human identification accuracy reaches 98.31% when incorporating matching clues, setting a high benchmark for model performance.

ii. Algorithmic Strength: Our algorithm surpasses others in analytical capabilities, achieving an accuracy 11% higher than GPT-4V with matching in non-human results. However, there remains a gap compared to human performance.

iii. Model Comparison: Among LLaVA-7b, LLaVA-13b, and GPT-4V with matching, GPT-4V exhibits the best performance with an accuracy of 68%, attributed to its superior analytical and reasoning abilities.

6.2 Group-VEQA Queries

Group-**VEQA** queries focus on identifying all pertinent clues for more reliable reasoning. The result is shown in Table 3.

Our method achieves the highest recall rate at 70.85%, outperforming GPT-4V, LLaVA-7b, and LLaVA-13b combined with matching by 5.81%, 30.81%, and 48.79%, respectively. This indicates that our approach excels in retrieval tasks compared to **MLLMs**, likely due to the effectiveness of rule-based methods in managing excessive information. Additionally, the performance of baseline **MLLMs** diminishes with reduced parameter sizes, suggesting a positive correlation between their analytical reasoning abilities and parameter sizes.

6.3 The Influence of Multi-Source Info

In principle, the effective recognition of personal information by a model depends on three main sources: its inherent knowledge, clues from the query, and clues from retrieved data. Our FRAG framework leverages these sources to guide accurate answers. As demonstrated in Table 4, when recall is accurate, LLaVA-7b correctly answers 42.86% of cases post-FRAG, while LLaVA-13b achieves 39.18%.

However, in practice, the presence of noise in the recalled information and the potential inability of **MLLMs** to effectively integrate FRAG information with the model's original knowledge may lead to incorrect answers. As shown in Table 4, LLaVA-7b+FRAG and LLaVA-13b+FRAG respectively provide incorrect answers in 7.32% and 9.44% of cases that could have been answered correctly before FRAG.

To assess the impact of the prompt on the model's original knowledge, we conducted ablation experiments by removing the Original-knowledge-aware Prompt (OP), as shown in Table 5. The

Table 6: Result for Singe-**VEQA** Queries of Common and Uncommon Entities.

Models	Common Entity $Acc^{hit}(\%)$	Uncommon Entity $Acc^{hit}(\%)$
LLaVA-7b	43.04	11.63
LLaVA-7b + FRAG	66.72	59.44
LLaVA-13b	51.60	14.34
LLaVA-13b + FRAG	66.38	59.09
GPT-4V	-	-
GPT-4V + FRAG	72.43	63.46
MAR	**81.24**	**77.19**

Table 7: Names Extracted from Original News in the **NewsPersonQA** Dataset and Their Frequencies

Name	Occurrence Frequency
Trump	3818
Obama	2737
Hillary Clinton	935
...	
Roger Clinton	4
Wayne Simmons	4

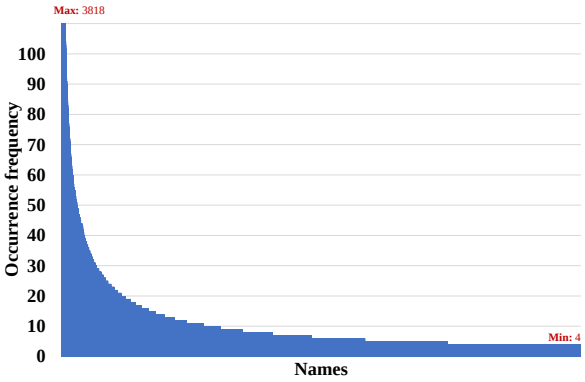

Figure 3: Diagram of names extracted from the original news in the **NewsPersonQA** dataset and their frequency of occurrence.

accuracy of LLaVA-7b, LLaVA-13b, and GPT-4V combined with FRAG decreased by 6.05%, 1.72%, and 4.51% respectively. These results highlight the importance of the model's own knowledge as a crucial clue in the reasoning process and underscore its significance in achieving accurate outcomes.

6.4 Analysis of Experimental Results for Common and Uncommon Entities

1. Name Distribution. We have tallied the frequency of names that appear four times or more in the original news files of the **NewsPersonQA** dataset. As shown in Table 7 and Fig. 3, it is evident that the dataset contains head-torso-tail entities, with torso-tail entities being less recognizable. We define head entities as those with a frequency greater than 50, which are mostly names of famous people; torso entities are those with a frequency between 10 and 50, representing a portion of the dataset; and tail entities are those with a frequency less than 10, which make up more than half of the entire dataset.

2. Experimental Results. We further conducted statistical analysis and evaluation on the experimental results presented in Section 6.1, specifically focusing on the results for common and uncommon entities (as shown in Table 6). Firstly, the performance of LLaVA-7b and LLaVA-13b indicates that **MLLMs** have a stronger recognition ability for common entities, but are less recognizable for torso-tail entities.

Secondly, with the addition of fine-grained RAG, LLaVA-7b and 13b showed an improvement of 23.68% and 14.78%, respectively, for common entities; and an improvement of 47.81% and 44.75% for uncommon entities. For GPT-4V, the addition of FRAG enabled it to respond to person entities, and due to its more powerful recognition and reasoning abilities, it achieved higher accuracy than LLaVa. However, by comparison, our method **MAR** demonstrated optimal performance in detecting both common and uncommon entities.

7 Conclusion

In this paper, we explore a novel visual-based (personal) entity questions (**VEQA**) problem that focuses on aggregating clues from multiple captioned visual objects. We introduce matching graphs designed to capture the relationships between identical entities across various visual objects. Extensive experiments demonstrate the high accuracy of our method. While our work has primarily focused on matching person entities, future research can aim to extend matching-augmented reasoning to other tasks.

Limitations

Currently, our framework primarily relies on similarity for face matching and does not consider factors such as age-related changes and facial blurring. This may result in inaccuracies in matching certain nodes, representing a future research direction. Additionally, in real-world applications, news is dynamic. Efficient retrieval and expansion strategies for a growing data lake pose challenges as the dataset evolves, warranting further investigation.

Ethics Statement

The authors declare that they have no conflict of interest. Our work aims to enhance the answer generation of visual question answering by retrieving entity-related clues. While improving the accuracy of answer generation, our method significantly saves resources as it does not require fine-tuning of large language models. We strive to ensure that our approach is not only accurate and efficient but also fair and unbiased. We recognize the potential of significant impact of visual question answering technology on society and pledge to maintain transparency in sharing our findings and progress with relevant users and stakeholders.

Acknowledgements

This paper is supported by NSF of China (62402409), Guangdong Basic and Applied Basic Research Foundation (2023A1515110545), and CCF-Huawei Populus Grove Fund (CCF-HuaweiDB202403).

References

Mayank Agrawal, Anand Singh Jalal, and Himanshu Sharma. 2023. A review on vqa: Methods, tools and datasets. In *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, pages 1–6. IEEE.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023a. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Zui Chen, Zihui Gu, Lei Cao, Ju Fan, Samuel Madden, and Nan Tang. 2023b. Symphony: Towards natural language query answering over multi-modal data lakes. In *13th Conference on Innovative Data Systems Research, CIDR 2023, Amsterdam, The Netherlands, January 8-11, 2023*. www.cidrdb.org.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq R. Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed representations of tuples for entity resolution. *Proc. VLDB Endow.*, 11(11):1454–1467.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

Ilija Ilievski and Jiashi Feng. 2017. Multimodal learning and reasoning for visual question answering. *Advances in neural information processing systems*, 30.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. Mm-reasoner: A multimodal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. 2024a. The dawn of natural language to SQL: are we fully ready? *Proc. VLDB Endow.*, 17(11):3318–3331.

Jiaqi Li, Miaozeng Du, Chuanyi Zhang, Yongrui Chen, Nan Hu, Guilin Qi, Haiyun Jiang, Siyuan Cheng, and Bozhong Tian. 2024b. Mike: A new benchmark for fine-grained multimodal entity knowledge editing. *arXiv preprint arXiv:2402.14835*.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. 2023b. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.

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

Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuyu Luo, Yuxin Zhang, Ju Fan, Guoliang Li, and Nan Tang. 2024a. A survey of NL2SQL with large language models: Where are we, and where are we going? *CoRR*, abs/2408.05109.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024b. Rar: Retrieving and ranking augmented mllms for visual recognition. *arXiv preprint arXiv:2403.13805*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

TR Mahesh, T Rajan, K Vanitha, HK Shashikala, et al. 2023. Intelligent systems for medical diagnostics with the detection of diabetic retinopathy at reduced entropy. In *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, pages 1–8. IEEE.

Weixing Mai, Zhengxuan Zhang, Kuntao Li, Yun Xue, and Fenghuan Li. 2023. Dynamic graph construction framework for multimodal named entity recognition in social media. *IEEE Transactions on Computational Social Systems*.

Xuedi Qin, Chengliang Chai, Nan Tang, Jian Li, Yuyu Luo, Guoliang Li, and Yaoyu Zhu. 2022. Synthesizing privacy preserving entity resolution datasets. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*, pages 2359–2371. IEEE.

Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Ethan Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, and Seungwhan Moon. 2024. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. *arXiv preprint arXiv:2403.04735*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. 2022. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in vqa. *arXiv preprint arXiv:2211.07516*.

Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. 2018. Learning visual knowledge memory networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7736–7745.

Yushi Sun, Xin Hao, Kai Sun, Yifan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2024. Are large language models a good replacement of taxonomies? *Proc. VLDB Endow.*, 17(11):2919–2932.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

Nan Tang, Chenyu Yang, Ju Fan, Lei Cao, Yuyu Luo, and Alon Y. Halevy. 2024. Verifai: Verified generative AI. In *14th Conference on Innovative Data Systems Research, CIDR 2024, Chaminade, HI, USA, January 14-17, 2024*. www.cidrdb.org.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Guoliang Li, Xiaoyong Du, Xiaofeng Jia, and Song Gao. 2023. Unicorn: A unified multi-tasking model for supporting matching tasks in data integration. *Proc. ACM Manag. Data*, 1(1):84:1–84:26.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. In *EMNLP (Findings)*. Association for Computational Linguistics.

Yupeng Xie, Yuyu Luo, Guoliang Li, and Nan Tang. 2024. Haichart: Human and ai paired visualization system. *arXiv preprint arXiv:2406.11033*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. 2024. CRAG - comprehensive RAG benchmark. *CoRR*, abs/2406.04744.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*.

Dongxiang Zhang, Rui Cao, and Sai Wu. 2019. Information fusion in visual question answering: A survey. *Information Fusion*, 52:268–280.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023a. Gpt-4v(ision) as a generalist evaluator for vision-language tasks.

Zhengxuan Zhang, Weixing Mai, Haoliang Xiong, Chuhan Wu, and Yun Xue. 2023b. A token-wise graph-based framework for multimodal named entity recognition. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2153–2158. IEEE.

Wenfeng Zheng, Yu Zhou, Shan Liu, Jiawei Tian, Bo Yang, and Lirong Yin. 2022. A deep fusion matching network semantic reasoning model. *Applied Sciences*, 12(7):3416.

Jie Zhu, Shufang Wu, Xizhao Wang, Guoqing Yang, and Liyan Ma. 2018. Multi-image matching for object recognition. *IET Computer Vision*, 12(3):350–356.

A Experimental Details

1. Setup and Environment: The experiments were conducted in a zero-shot setting using RTX 4090 GPUs, with PyTorch version 1.12.0. For GPT-4V, we used the interface of the GPT-4-vision-preview model. It is worth noting that GPT-4V often refrains from answering person identification questions without additional clues due to policy reasons. However, with the incorporation of matching graph techniques, it can leverage weak signals and combine them with its own knowledge base.

2. Efficiency and Time: For preprocessing, using DeepFace for face detection and extraction from an image takes approximately 0.1 to 0.4 seconds. Performing NER on captions using spaCy takes about 0.001 seconds per caption. Additionally, processing each query, which includes retrieval, constructing a matching graph for the query, and reasoning, takes 0.01 to 0.3 seconds to complete the entire process.

3. Parameters: We determined the experimental hyperparameters by creating a small sample of approximately 100 data points. During node retrieval, the face similarity threshold σ_f and name similarity threshold σ_n were set to 0.8 and 0.9, respectively. The number of iterations k for node retrieval was set to 2, and the maximum number of seed nodes was set to 10. It is worth noting that variations in these hyperparameters have little impact on the experimental results, as **MLLMs** can correctly answer questions when the hit includes correct examples. Thus, our method still demonstrates strong generalizability.