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

Anonymous ACL submission

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 methodology 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 this information and stores their vector representations in vector databases. When handling **VEQA**, **MAR** retrieves matching faces and names and organizes these entities into a matching graph, where nodes represent entities and edges indicate their similarities. **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 Introduction

Multimodal language models (**MLLMs**) (Cui et al., 2024) like GPT-4V (Zhang et al., 2023) 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, especially for personal entities.

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**.

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., 2024). 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). 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.

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, 2024; Yu et al., 2023).

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. Note, in this study, our focus is on personal entities. We plan to extend our analysis to include additional types of entities in future research.

As illustrated in Figure 1(c), if we can successfully 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 easily answer Q_2.

Contributions. Our notable contributions are summarized as follows.

- We study **VEQA**, an important and commonly used subset of VQA, but is under-explored. (Section 3)
- 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 effective answer a **VEQA**. (Section 4)
- Given that **VEQA** is a relatively new problem, existing benchmarks are not suitable. Therefore, we have constructed a new benchmark **NewsPersonQA** including 235k images and 6k QA pairs. (Section 5)
- We conduct extensive experiments to show that **MAR** $>$ **MLLMs** + **RAG** $>$ **MLLMs**, where **RAG** is to feed the retrieved matching graph to **MLLMs**. (Section 6)

2 Related Work

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.

MLLMs for VQA. **MLLMs**, such as GPT-4V (Zhang et al., 2023) 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.

RAG for VQA. However, in many cases, the cues within images and text are insufficient for reasoning and answering. Retrieval-augmented generation (RAG) (Lewis et al., 2021) 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; Lin et al., 2022).

VEQA. In this paper, we investigate **`VEQA`**, a critical subset of VQA that concentrates on querying information about entities, especially persons. As will be shown in Section 6, **`MLLMs`** often struggle with such questions due to limited knowledge and privacy considerations. While RAG can enhance **`MLLMs`** for **`VEQA`** tasks, **`MLLMs`** still face challenges (or confused) in reasoning with multiple interconnected visual objects.

Data Matching. Data matching refers to the process of identifying, comparing, and merging records from multiple datasets to determine whether they correspond to the same entities (Christen and Christen, 2012). With the increasing multimodality of data, the concept of matching has been continually expanded from its original string matching (Text-Text) and entity matching (Tuple-Tuple) context. For instance, Image-Text Matching (Lee et al., 2018; Li et al., 2019), Image-Image matching (Zhu et al., 2018), etc. In fact, matching can aggregate more clues, enhance the reasoning ability of models, and possess strong interpretability (Zheng et al., 2022).

3 Problem

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.

Users can pose a Visual-based (Personal) Entity Question Answering (**`VEQA`**) on either a single captioned visual object (Single-**`VEQA`**) or a group of such objects (Group-**`VEQA`**), as defined below.

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” or “is he/she John”.

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 $\mathbf{Q}$ to represent either a Single-**`VEQA`** or a Group-**`VEQA`** query, when it is clear form the context.

4 Algorithms for VEQA

In this section, we will first discuss solely using **`MLLMs`** for **`VEQA`** in Section 4.1. We will then discuss coarse-grained retrieval-augmented generation (RAG) in Section 4.2. We then propose a new concept, called matching graphs, which can provide fine-grained information among retrieved objects in Section 4.3. Based on matching graphs, 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:

$$\boxed{\mathtt{Q} \rightarrow \mathtt{MLLM} \rightarrow \text{answer}}$$

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:

$$\boxed{(\mathtt{Q}, \text{top-}k\text{ objects}) \rightarrow \mathtt{MLLM} \rightarrow \text{answer}}$$

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.

If we are certain about a person's name, we will use a square bracket *e.g.,* **name**(n) = [Yi Wang] for the selected face in Figure 1(a); if we are not sure about a person's name, we will

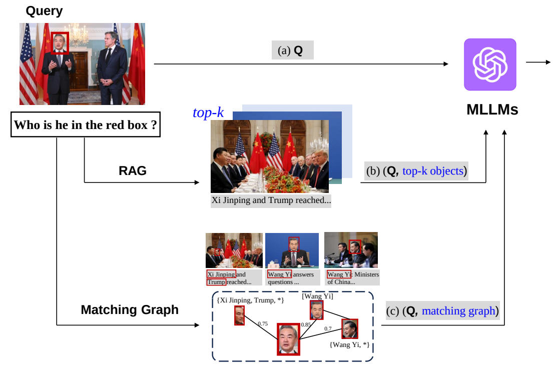

Figure 2: Different algorithms for **VEQA**. (a) **MLLMs**. (b) Coarse-grained RAG. (c) Fine-grained RAG.

use a curly bracket to indicate possible names *e.g.,* **name**(n) = {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.,* **face**(n_i)) and n_j (*i.e.,* **face**(n_j)) are likely to be the same person. Each edge has a weight **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) $\rightarrow$ **MLLM** $\rightarrow$ 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 **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**, **V**$'$, and **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.

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.

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**

Table 1: Statistics of **NewsPersonQA**

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**.

6 Experiment

Methods. For answering **VEQA** queries, we selected two well-known and highly capable **MLLMs** 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.

Models	Acc (%)	$\textbf{Acc}^{\texttt{hit}}$ (%)
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 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.)

- 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.

+ **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 $\textbf{Acc}^{\texttt{hit}}$. For Group-**`VEQA`** queries, we employ recall (**Recall**) as the metric.

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-

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

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

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 Further Study - The Influence of Multi-Source Information

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

Models	Acc^{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 4: Study on Successfully Recalled Data.

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 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.

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%

Table 5: Original-knowledge-aware Prompt (OP) ablation study result

7 Conclusion

In this paper, we explore a novel **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.

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*.

Peter Christen and Peter Christen. 2012. *The data matching process*. Springer.

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.

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

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 multi-modal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216.

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.

Jiaqi Li, Miaozeng Du, Chuanyi Zhang, Yongrui Chen, Nan Hu, Guilin Qi, Haiyun Jiang, Siyuan Cheng, and Bozhong Tian. 2024. 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*.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

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

Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. 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.

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.

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.

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.

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*.

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. 2023. Gpt-4v(ision) as a generalist evaluator for vision-language tasks.

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.