Enhancing Visual Question Answering through Question-Driven Image Captions as Prompts

Övgü Özdemir, Erdem Akagündüz
Department of Modeling and Simulation,
Graduate School of Informatics,
Middle East Technical University, Türkiye
{ovgu.ozdemir,akaerdem}@metu.edu.tr

Abstract

Visual question answering (VQA) is known as an AI-complete task as it requires understanding, reasoning, and inferring about the vision and the language content. Over the past few years, numerous neural architectures have been suggested for the VQA problem. However, achieving success in zero-shot VQA remains a challenge due to its requirement for advanced generalization and reasoning skills. This study explores the impact of incorporating image captioning as an intermediary process within the VQA pipeline. Specifically, we explore the efficacy of utilizing image captions instead of images and leveraging large language models (LLMs) to establish a zero-shot setting. Since image captioning is the most crucial step in this process, we compare the impact of state-of-the-art image captioning models on VQA performance across various question types in terms of structure and semantics. We propose a straightforward and efficient question-driven image captioning approach within this pipeline to transfer contextual information into the question-answering (QA) model. This method involves extracting keywords from the question, generating a caption for each image-question pair using the keywords, and incorporating the question-driven caption into the LLM prompt. We evaluate the efficacy of using general-purpose and question-driven image captions in the VQA pipeline. Our study highlights the potential of employing image captions and harnessing the capabilities of LLMs to achieve competitive performance on GQA under the zero-shot setting. Our code is available at https://github.com/ovguyo/captions-in-VQA.

1. Introduction

Visual Question Answering (VQA) is a complex multi-modal task that demands a high-level understanding of several aspects, such as object and attribute identification, object localization, comprehension of the relationship between the image and the question, and reasoning about the context and the scene. The common steps of a typical VQA model involve generating embeddings of the image and the question using encoders for each, combining the image and question embeddings with a fusing module, and generating answers using a text generator or a classifier. For a general overview of the VQA techniques, the reader may refer to [33, 34].

The inherent multimodal nature of the VQA problem is the primary factor contributing to its complexity. Combining different types of information, such as text and images, makes the model’s training more complex, as the model must understand and utilize the connections and interactions between these different modalities. Several studies [4, 15, 18, 28, 32] propose an approach to tackle multimodality for the VQA problem. However, these methods indicate limitations in their capacity to adapt to new tasks, particularly in zero-shot settings.

Recent advances in high-capacity large language models (LLMs) [1, 5, 36] have marked a dramatic milestone in the domain. LLMs are predominantly trained with millions (or billions) of parameters and utilized for processing textual data. LLMs show outstanding performance in a variety of natural language tasks. The ongoing research challenge lies in extending the capabilities of LLMs to the intersection of different modalities, e.g., textual and visual data. Recently, GPT-4 [1] and Gemini [36] stand out as remarkable examples of multimodal LLMs, adept at successfully processing textual and visual modalities for various downstream tasks, including VQA. Several alternative approaches [2, 7, 19, 20, 22] have also been proposed in the realm of large-scale vision-language integration. The challenge in multimodal training lies in the extensive computational and data costs required to align the representation spaces of vision and language.

Some recent studies [11, 37, 42] delve into the poten-
tial of utilizing image captions with unimodal LLMs in the zero-shot VQA setting. Our study differs from these studies in the following aspects. Firstly, we focus on examining the representation capacity of image captions from various vision-language models on the VQA performance. Second, our study investigates whether image captions can be informative for specific types of questions by evaluating the results in structurally and semantically different questions. Within this scope, we also evaluate the influence of feeding LLMs with general-purpose and question-driven captions, and only the most relevant sentence in the caption during the QA stage.

Numerous VQA datasets are available in the literature, including CLEVR [17], VQA [3], VQA 2.0 [9], OK-VQA [25], GQA [16]. Among these sets, although each serves various purposes effectively, GQA stands out for its emphasis on testing compositional and grounded reasoning abilities and its relatively diverse Q/A set. In this study, we conduct our experiments on the GQA dataset and focus on measuring performance on semantically and structurally different questions.

We structure the VQA task into two fundamental components: image captioning and question-answering. The goal is to leverage the respective strengths of these tasks, aiming for a more thorough comprehension of both the visual content and the corresponding questions. We carry out experiments with state-of-the-art vision-language models, including CogVLM [40], BLIP-2 [20], and FuseCap [31] to comprehend their scene representation capacity in the VQA pipeline.

We outline our contributions as follows:

- We evaluate the image captioning performance of various vision-language models incorporating them with LLMs for zero-shot VQA, analyzing their effectiveness across various question types.
- We propose a straightforward question-driven captioning approach to better transfer the context into LLMs for question-answering.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 mentions the components of the proposed pipeline. In Section 4, we present the experiments designed for our study. Section 5 discusses evaluation results. Section 6 outlines the conclusions drawn from our findings and discusses potential avenues for future research.

2. Related Literature

2.1. Large Language Models

LLMs [5, 27, 38] trained on extensively rich web-scale corpus usually employ autoregressive methods to generate target tokens. LLMs demonstrate remarkable proficiency in processing and generating text with human-like characteristics. This attribute renders them suitable instruments for various language-related tasks, including question-answering, text generation, machine translation, etc. Expanding the scope of LLMs to include additional modalities results in the creation of multimodal LLMs [1, 20, 22, 36, 40], which boosts the performance for many downstream tasks including image captioning, visual question answering, text-to-image synthesis.

2.2. Visual Question Answering

The main challenge of the VQA domain comes from bridging the gap between visual understanding and natural language. Numerous studies have been proposed to tackle questions related to visual content. Relation Networks [32] involves employing a compact and straightforward neural network module that takes pairs of features as input and generates a score indicative of the relationship between these feature pairs. LXMERT [35] is a large-scale transformer model that fuses textual and visual representations with a cross-modality encoder. MDETR [18] is an end-to-end modulated detector which is an improved version of the object detection model DETR [6] by adding the capability of processing free-form texts. Alternatively, neuro-symbolic approaches in VQA have gained attention to enhance model interpretability. A neuro-symbolic approach in VQA combines two main parts: neural network modules for handling images and text modalities, and a symbolic reasoning module for managing logic and knowledge representation. NS-VQA [43] and NS-CL [24] use neural networks for scene parsing and dividing questions into program instructions, and propose a symbolic module executing the program instructions on the scene representation. An alternative hybrid approach, ProTo [44], proposes program-guided transformers that use semantic and structural information of the programs being parsed from the questions by a sequence-to-sequence model. A recent approach, namely VisProg [12], generates program instructions from questions using LLMs and employs instructions on images benefiting from different modules for object detection, visual question answering, image classification, and segmentation. Recent large-scale multimodal approaches used for VQA are mentioned in Section 1 and Section 2.1.

2.3. Image Captioning

Image captioning aims to produce a caption describing visual content in natural language. Conventional approaches in image captioning are based on attention and encoder-decoder structure [13, 14, 41]. A typical image captioning model consists of an encoder for gathering visual cues and a textual decoder to produce the final caption. Like VQA, this requires bridging the gap between visual and natural language understanding. Recently, large-scale multimodal models [1, 12, 19, 20, 26, 36, 40] have resulted
in notable enhancements in performance and demonstrated adaptability to various downstream applications, including image captioning.

2.4. Question Answering

Question-answering (QA) models aim to provide contextually appropriate responses based on a document or text, often requiring an understanding of linguistic rules, syntax, and contextual nuances. Recent models in QA leverage transformer architectures and large-scale pre-training on diverse datasets [5, 8, 23, 29].

3. Methodology

3.1. Caption Generation

The primary and most crucial element in the suggested pipeline is the creation of image captions with high visual representation capability. Image captions provide a summarized version of the visual content, and specific visual details may be lost, which could affect the VQA performance. We survey image captioning models, selecting ones that provide more detailed captions while taking into account our computational resource limitations. Consequently, we evaluate several zero-shot vision-language models, including CogVLM [40], FuseCap [31], and BLIP-2 OPT$_{2,7b}$ [20] by integrating them into the VQA pipeline. We employ both the chat and visual grounding variants of CogVLM, considering their potential performance impacts across different question types. VQA performance is assessed across various image captions according to structurally and semantically different question categories. More details about question categories are given in Section 4.1.

Two approaches are utilized in this paper to generate captions. First, each image is captioned without considering the questions associated with it, which we refer to as “general-purpose captioning” throughout the paper. However, general-purpose captions are designed to provide a broad description of the visual content, and they may lack the precision needed to address detailed and specific queries. Therefore, in our second approach, we create image captions for each image-question pair, a process we refer to as “question-driven image captioning”. For this purpose, KeyBERT [10] is employed to extract keywords from the questions. KeyBERT utilizes BERT-embeddings along with a basic cosine similarity measure to identify the most representative words that encapsulate the content of the entire text. Extracted keywords are fed into the image captioning model along with the corresponding image, as illustrated in Figure 1.

We also investigate whether less relevant portions of an image caption could potentially introduce confusion or result in inaccurate answers for the QA model/LLM. Hence, in our analysis, we experiment with keeping only the most relevant sentence of the image caption and providing it to the LLM during the QA step. To achieve this, we utilize Sentence-BERT [30], specifically employing the MiniLM-L6 model$^1$, to extract the most relevant sentence from the image caption based on the given question.

3.2. Question Answering

As in the pipeline shown in Figure 1, the QA model takes the image caption and the question as input, leveraging information from the image caption to generate an answer. During the QA step, we utilize GPT-3.5, recognized for its high zero-shot performance in QA benchmarks [39]. Despite the superior performance of the more recent LLMs, GPT-4, across various natural language tasks including QA, we choose not to use GPT-4 in our experiments to keep our pipeline cost-effective. In future works, the integration of higher-performing LLMs with the pipeline could be explored.

We derive answers with an open-ended generation, specifically using GPT-3.5-turbo API provided by OpenAI. The answer size is restricted to a maximum of two words, aligning with the answer size distribution in the GQA dataset. Optimal prompts are given in the Section 4.5.

4. Experimental Setup

4.1. Dataset

We conduct experiments on the GQA [16] dataset, specifically the balanced version of the test-dev subset, comprising 12,578 questions. Each image in the dataset is linked to multiple questions, and the overall number of images included is 398. This subset contains a diverse distribution of questions across various categories, with a primary focus on categorization based on structure and semantics. The structural type is determined by the final operation in the functional program of the question, encompassing categories such as verify, query, choose, logical, and compare. The semantic type specifies the primary focus of the question and includes categories like object, attribute, category, relation, and global. Table 1 presents an overview of question types, corresponding descriptions, and the respective number of questions in the GQA test-dev [16].

4.2. Competing VQA Methods

To evaluate zero-shot VQA performance, we use the chat variation of CogVLM$^2$ and BLIP-2 FlanT5$_{X.L}$$^3$. CogVLM is an open-sourced pre-trained vision-language model with 10B visual and 7B language parameters. CogVLM outperforms many vision-language models, e.g., InstructBLIP

$^1$https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1
$^2$https://huggingface.co/THUDM/cogvlm-chat-hf
$^3$https://huggingface.co/Salesforce/blip2-flan-t5-xl
Figure 1. VQA pipeline exploiting general and the proposed question-driven (QD) image captioning as an intermediate step.

Table 1. Overview of the question types

<table>
<thead>
<tr>
<th>Question type</th>
<th>Description</th>
<th>Example</th>
<th>No. samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>verify</td>
<td>yes/no questions</td>
<td>Does the device under the picture frame look black?</td>
<td>2252</td>
</tr>
<tr>
<td>query</td>
<td>open questions</td>
<td>Which kind of vehicle is waiting for the traffic light?</td>
<td>6805</td>
</tr>
<tr>
<td>choose</td>
<td>choosing from alternatives</td>
<td>What color is the hair, gray or red?</td>
<td>1128</td>
</tr>
<tr>
<td>logical</td>
<td>logical inference</td>
<td>Are the flags triangular and red?</td>
<td>1803</td>
</tr>
<tr>
<td>compare</td>
<td>comparison of objects</td>
<td>Which is larger, the pasture or the horse?</td>
<td>589</td>
</tr>
<tr>
<td>object</td>
<td>existence questions</td>
<td>Are there both a horse and a fence in the image?</td>
<td>778</td>
</tr>
<tr>
<td>attribute</td>
<td>object properties/position</td>
<td>On which side of the picture are the pens?</td>
<td>5185</td>
</tr>
<tr>
<td>category</td>
<td>object identification</td>
<td>What kind of clothing is yellow?</td>
<td>1149</td>
</tr>
<tr>
<td>relation</td>
<td>relations with objects/subjects</td>
<td>Is the toaster to the right of a refrigerator?</td>
<td>5308</td>
</tr>
<tr>
<td>global</td>
<td>overall properties</td>
<td>Is it an outdoors scene?</td>
<td>157</td>
</tr>
</tbody>
</table>

[7] and LLaVA-1.5 [21], in VQA benchmarks. Due to our resource constraints with 16 GB VRAM, we apply 4-bit quantization to CogVLM, BLIP-2 FlanT5\_XL with 4.1B parameters also indicate high performance surpassing BLIP-2 OPT\_6.7B and Flamingo [2] in VQA benchmarks. We employ BLIP-2 FlanT5\_XL with F16 precision.

4.3. Image Captioning Methods

We examine the VQA performance attributed to semantic and structural question types mentioned in Section 4.1. Image captions are obtained through the visual grounding\(^4\) and chat\(^5\) variations of the CogVLM, FuseCap\(^6\), and BLIP-2 OPT\_2.7\_b\(^7\) models. When determining the image captioning method, we pay attention to both its alignment with our resource capacity and its high performance in image captioning benchmarks. We employ 4-bit quantization to CogVLM and use F16 precision for BLIP-2 OPT\_2.7\_b.

4.4. Evaluation

Before the evaluation, GPT-3.5 predictions undergo post-processing, which involves the removal of punctuation. During the evaluation process, we employ the accuracy metric, calculated as the ratio of correctly predicted answers to the total number of answers. Given that answers are derived through open-ended generation using LLMs and might include variations, we do not seek an exact match between the prediction and the ground truth. Instead, we evaluate semantic similarity using cosine similarity in a vector space with the threshold $0.70$. If two strings are closely aligned in meaning, the prediction is accepted as correct; for example, accepting couch as correct for the label sofa. We determine the similarity threshold through manual observation of the results. At lower thresholds, we observe that predictions incorporating words related to each other, yet lacking identical meanings, are also considered correct. For instance, the similarity value between the words blue and brown is found to be $0.67$. We additionally assess performance across higher cosine similarity thresholds, e.g. 0.8 and 0.9, and for exact matching (EM).

4.5. Prompt Details

A brief prompt, ‘Describe the scene in this image’ is supplied to the image captioning model to create general-purpose image captions. To create question-driven captions, ‘Consider the keywords: [keywords]’ is added to the prompt. In the QA stage, the LLM prompt involves ‘Answer the question in a maximum of two words based on the
Table 2. Comparison of the performances of different image captioning methods in the context of VQA on GQA test-dev. Image captioning methods are employed with GPT-3.5 as the question-answering (QA) method. Two variants of CogVLM, namely visual grounding (CogVLM-V) and chat model (CogVLM-C), are utilized for image captioning. QD and SB refer to question-driven and sentence-based captions, respectively. The answers with a cosine similarity of 0.7 or higher have been considered correct with the label. Accuracy values are compared with the performance of zero-shot VQA models based on various question categories.

<table>
<thead>
<tr>
<th>Question type</th>
<th>CogVLM-C Cap. + GPT-3.5 QA</th>
<th>CogVLM-V Cap. + GPT-3.5 QA</th>
<th>CogVLM-C QD Cap. + GPT-3.5 QA</th>
<th>CogVLM-C SB Cap. + GPT-3.5 QA</th>
<th>FuseCap Cap. + GPT-3.5 QA</th>
<th>BLIP-2 Cap. + GPT-3.5 QA</th>
<th>CogVLM VQA</th>
<th>BLIP-2 VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>verify</td>
<td>63.01</td>
<td>58.53</td>
<td>66.83</td>
<td>61.06</td>
<td>53.60</td>
<td>55.82</td>
<td>83.04</td>
<td>56.48</td>
</tr>
<tr>
<td>query</td>
<td>36.91</td>
<td>31.08</td>
<td>38.34</td>
<td>31.51</td>
<td>29.61</td>
<td>31.87</td>
<td>54.11</td>
<td>41.31</td>
</tr>
<tr>
<td>choose</td>
<td>65.25</td>
<td>60.90</td>
<td>65.51</td>
<td>60.90</td>
<td>58.07</td>
<td>60.82</td>
<td>87.32</td>
<td>56.91</td>
</tr>
<tr>
<td>logical</td>
<td>59.51</td>
<td>60.29</td>
<td>59.07</td>
<td>58.46</td>
<td>57.07</td>
<td>56.07</td>
<td>77.54</td>
<td>54.24</td>
</tr>
<tr>
<td>compare</td>
<td>51.78</td>
<td>51.95</td>
<td>51.95</td>
<td>49.07</td>
<td>54.50</td>
<td>48.22</td>
<td>62.65</td>
<td>46.52</td>
</tr>
<tr>
<td>object</td>
<td>61.95</td>
<td>63.24</td>
<td>59.13</td>
<td>58.87</td>
<td>59.38</td>
<td>58.35</td>
<td>84.45</td>
<td>57.07</td>
</tr>
<tr>
<td>attribute</td>
<td>51.75</td>
<td>46.42</td>
<td>54.62</td>
<td>50.80</td>
<td>45.11</td>
<td>46.63</td>
<td>70.45</td>
<td>49.33</td>
</tr>
<tr>
<td>category</td>
<td>47.35</td>
<td>44.21</td>
<td>50.39</td>
<td>42.56</td>
<td>43.52</td>
<td>42.47</td>
<td>63.19</td>
<td>53.35</td>
</tr>
<tr>
<td>relation</td>
<td>42.56</td>
<td>38.32</td>
<td>42.97</td>
<td>35.76</td>
<td>34.98</td>
<td>37.23</td>
<td>59.91</td>
<td>43.31</td>
</tr>
<tr>
<td>global</td>
<td>49.04</td>
<td>45.86</td>
<td>45.86</td>
<td>44.59</td>
<td>43.95</td>
<td>45.22</td>
<td>56.05</td>
<td>40.13</td>
</tr>
<tr>
<td>total</td>
<td>48.06</td>
<td>43.83</td>
<td>49.50</td>
<td>44.12</td>
<td>41.58</td>
<td>42.99</td>
<td>66.02</td>
<td>47.52</td>
</tr>
</tbody>
</table>

5. Results

5.1. Main Findings

Table 2 summarizes our results and demonstrates that employing our suggested QD image captioning approach for VQA enhances performance across most question categories compared to general-purpose image captioning. Also, Table 3 indicates that the QD image captioning approach utilizing the CogVLM-chat variant surpasses other image captioning methods in evaluations seeking both different cosine similarity thresholds and exact matching.

Significant performance enhancements are evident in QD image captions, particularly in the verify category for yes/no questions, as well as attribute and category types primarily focused on identifying and describing a single object’s properties. However, challenges arise in the object category often asking which of two objects exists in the frame. Particularly in this category of questions, despite the QD image captions containing relevant information, inaccuracies emerge due to the behavior of the QA model, as elaborated in Section 5.2.

We also notice that the QD captioning emphasizing question keywords is linked to a performance decline in the global type questions. Global-type questions typically pertain to the overall content of an image. It suggests that the emphasis on question keywords in the caption negatively affects the model’s ability to make inferences about the entire image. On the other hand, it is quite possible to give other answers to questions of this type that are meaningful and contextually correct but do not match the label. In most of the cases, we observe that GPT-3.5 predicts answers that could be correct but do not precisely match the expected label (see examples in Figure 3).

In most question categories, the accuracy achieved by combining QD image captions with GPT-3.5 for VQA exceeds the performance of BLIP-2 FlanT5 XL in the zero-shot setting. However, all image captioning-based approaches indicate inferior performance compared to the CogVLM-chat model for VQA. We are intrigued to discover a notable disparity in performance when comparing

Table 3. Comparison of overall accuracy for exact matching (EM) and in different cosine similarity thresholds.

<table>
<thead>
<tr>
<th>Models</th>
<th>EM</th>
<th>sim=0.9</th>
<th>sim=0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>CogVLM-C Cap. + GPT-3.5 QA</td>
<td>36.77</td>
<td>38.21</td>
<td>43.01</td>
</tr>
<tr>
<td>CogVLM-V Cap. + GPT-3.5 QA</td>
<td>36.21</td>
<td>37.51</td>
<td>41.21</td>
</tr>
<tr>
<td>CogVLM-C QD Cap. + GPT-3.5 QA</td>
<td>37.64</td>
<td>39.24</td>
<td>44.48</td>
</tr>
<tr>
<td>CogVLM-C SB Cap. + GPT-3.5 QA</td>
<td>34.14</td>
<td>35.06</td>
<td>39.41</td>
</tr>
<tr>
<td>FuseCap Cap. + GPT-3.5 QA</td>
<td>33.17</td>
<td>34.18</td>
<td>37.64</td>
</tr>
<tr>
<td>BLIP-2 Cap. + GPT-3.5 QA</td>
<td>34.77</td>
<td>35.53</td>
<td>39.11</td>
</tr>
<tr>
<td>CogVLM VQA</td>
<td>58.43</td>
<td>59.23</td>
<td>62.79</td>
</tr>
<tr>
<td>BLIP-2 VQA</td>
<td>37.82</td>
<td>38.57</td>
<td>42.33</td>
</tr>
</tbody>
</table>
Figure 2. Examples from correct predictions in case that QD image captioning is applied.
Figure 3. Examples from wrong predictions in case that QD image captioning is applied.
the image captions extracted by the CogVLM-chat model and provided to LLM, in contrast to the VQA performance of the CogVLM-chat model, unlike the case with BLIP-2.

Among the FuseCap, BLIP-2 OPT, CogVLM-chat, and CogVLM-visual grounding models, the most informative captions for VQA are obtained through the CogVLM-chat variant. The CogVLM-visual grounding variant indicates the highest performance only in object and logical question categories. This suggests that visual grounding models may provide an advantage in these question categories with their capacity to connect language queries to relevant visual elements and reason about object-related relationships.

Limiting image captions to the most relevant sentence reduces the overall performance of the CogVLM-chat model, though the impact varies across question types, with verify, query, choose and relation types being more negatively affected. This suggests that sentences less directly related to the questions do not result in confusion or inaccuracies for LLM during the QA. Conversely, generating more comprehensive and context-rich image captions is necessary for optimal performance.

Figure 2 and 3 feature examples of both correct and incorrect outcomes, where image captions are generated by the CogVLM-chat model using question-driven captioning and then fed to GPT-3.5 for answer prediction.

5.2. Error Analysis

When examining incorrect predictions based on question types, we discover some common issues.

We notice that 27% of the incorrect predictions are related to yes/no questions. A closer look reveals that in 11% of the incorrectly answered yes/no questions, GPT-3.5 provides a response using a word other than yes or no. For instance, when the provided caption is ‘The image showcases a skateboarder in action, possibly performing a trick on a ramp. The skateboarder is wearing protective gear, including a helmet, knee pads, and elbow pads. The background features a clear blue sky, trees, and a building. The overall ambiance suggests an outdoor skateboarding event or practice session.’, in response to the question ‘Are there salt shakers or skateboards in the picture?’, GPT-3.5’s prediction is skateboards, while the ground-truth is yes. We observe that most similar inaccuracies are associated with questions related to object and logical types, often connecting more than one object or attribute using conjunctions like and or or, as given in the example. We posit that this issue can be alleviated by crafting more effective prompts for GPT-3.5 or by employing a more powerful LLM for QA.

We also assess the instances where the LLM fails to provide an answer based on the information present in the image caption. Specifically, we examine the occurrences of not mentioned and not visible responses from GPT-3.5. Our findings indicate that, for the best general-purpose image captioning model, GPT-3.5 is not able to respond to 1.7% of the questions. Notably, when employing question-driven captioning, this rate decreases to 0.5%.

6. Conclusion

This study aims to develop a zero-shot VQA pipeline, leveraging LLMs with the inclusion of image captioning as an intermediate step, and evaluate its performance on the GQA benchmark. The proposed approach involves question-driven image captioning to transfer contextual information to the QA model. The study includes a thorough evaluation of zero-shot models for image captioning in the VQA context, comparing the impact of general-purpose and question-driven image captions in terms of various types of questions. Our comparative analysis suggests that incorporating question-driven image captions into the VQA process has a more favorable effect on overall performance, surpassing the VQA performance of BLIP-2. Future endeavors may explore the integration of larger-scale LLMs, e.g., GPT-4, to further enhance performance. Additionally, evaluating the pipeline in a few-shot setting could offer a more comprehensive comparison. To enhance transparency, replacing the QA model with an interpretable alternative, such as graph-based QA models, can be explored.

Acknowledgements

This work is partially supported by Middle East Technical University Scientific Research Projects Coordination Unit (METU-BAP), under the project number ADEP-704-2024-11482.

References


