2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-6547-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPRW63382.2024.00163

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 multimodal 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-

2160-7516/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPRW63382.2024.00163

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 LLM, 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$_{XL}$[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.

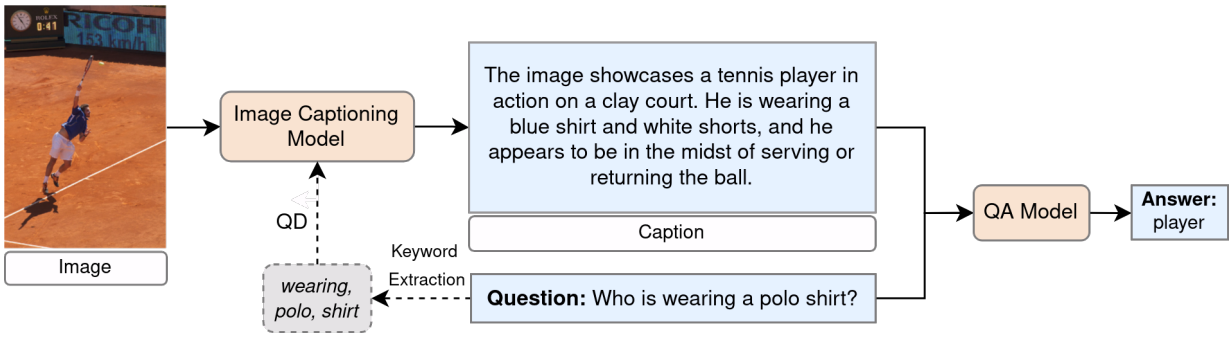

Table 1. Overview of the question types

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

[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$_{2.7b}$[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$_{2.7b}$.

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

[4]https://huggingface.co/THUDM/cogvlm-grounding-generalist-hf
[5]https://huggingface.co/THUDM/cogvlm-chat-hf
[6]https://github.com/RotsteinNoam/FuseCap
[7]https://huggingface.co/Salesforce/blip2-opt-2.7b

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.

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

text. Consider the type of question in your answer. For example, if it is a yes/no question, the answer should be yes or no. Text: `[text]`, Question: `[question]`'. We notice a positive impact on the results when we include an instruction in the prompt to consider the question type. In the decoding step for the answer generation, we set $temperature$ as 0.2, top_p as 1, and specify $frequency\_penalty$ and $presence\_penalty$ as 0.

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).

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

Models	EM	sim=0.9	sim=0.8
CogVLM-C Cap. + GPT-3.5 QA	36.77	38.21	43.01
CogVLM-V Cap. + GPT-3.5 QA	36.21	37.51	41.21
CogVLM-C QD Cap. + GPT-3.5 QA	**37.64**	**39.24**	**44.48**
CogVLM-C SB Cap. + GPT-3.5 QA	34.14	35.06	39.41
FuseCap Cap. + GPT-3.5 QA	33.17	34.18	37.64
BLIP-2 Cap. + GPT-3.5 QA	34.77	35.53	39.11
CogVLM VQA	**58.43**	**59.23**	**62.79**
BLIP-2 VQA	37.82	38.57	42.33

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

Question: How do the cars look like, dense or sparse?
Label: dense
Prediction: dense
Semantic_type: attr
Structural_type: choose
Text: The scene in the image can be described as 'dense' with 'cars' being a prominent element. The banners and signs add a 'like' element to the urban setting, making it look 'sparse' in comparison to the dense arrangement of vehicles.

Question: Is the river wide or is it narrow?
Label: narrow
Prediction: narrow
Semantic_type: attr
Structural_type: choose
Text: The image showcases a bridge spanning over a narrow river, surrounded by lush greenery and dotted with birds in flight. The sky above is vast and filled with clouds, suggesting a dynamic weather pattern.

Question: Which side of the picture is the plastic container on, the right or the left?
Label: right
Prediction: right
Semantic_type: attr
Structural_type: choose
Text: The scene in the image depicts a woman, possibly in her kitchen, reaching into an open refrigerator. She is wearing a white tank top and blue jeans. The refrigerator contains various food items, some of which are in containers and plastic packaging. The image seems to capture a candid moment of the woman searching for something in the fridge.

Question: Do the mountain side and the pole have the same color?
Label: no
Prediction: no
Semantic_type: attr
Structural_type: compare
Text: The image showcases a snowy mountain landscape where two individuals are skiing. One person is actively skiing downhill, while the other is seated on the snow, possibly taking a break or adjusting their equipment. The predominant colors in the scene are white (from the snow), green (from the trees), and red (from the clothing of the skiers).

Question: Is the hat the same color as the uniform?
Label: yes
Prediction: yes
Semantic_type: attr
Structural_type: compare
Text: The image showcases a young baseball player in a blue uniform, wearing a matching blue hat. The player is in the act of throwing the ball, with a focused expression on his face. The uniform is complemented by white socks with red stripes. The background consists of a grassy field with a chain-link fence, and there are trees visible in the distance.

Question: Are both the helmet and the bat made of the same material?
Label: no
Prediction: no
Semantic_type: attr
Structural_type: compare
Text: The scene depicts a baseball game where a player is in the midst of a swing, using a wooden bat. The player is wearing a helmet for protection, and the bat appears to be made of wood.

Question: What are the drapes around of?
Label: window
Prediction: window
Semantic_type: rel
Structural_type: query
Text: The image showcases a cozy room with white drapes cascading from a window. The room is adorned with various personal items, including a bed with blue bedding, a green sofa, a wooden chest, and a small table. There's also a fireplace with a decorative piece on top and a mirror hanging above it. The room exudes a sense of comfort and personal touch.

Question: Who is wearing a shirt?
Label: girl
Prediction: woman
Semantic_type: rel
Structural_type: query
Text: The image showcases a woman wearing a turquoise shirt while walking on a sandy beach. She is holding a surfboard under her arm and has tattoos on her legs. In the background, there are palm trees, a hut, and some boats, suggesting a tropical beach setting.

Question: Is the bag made of leather lying on top of a sofa?
Label: no
Prediction: no
Semantic_type: rel
Structural_type: verify
Text: The image showcases various personal items scattered on a white bedspread. Prominent objects include a green leather handbag, a mobile phone, a notebook, and several other miscellaneous items. The items are laid out in a haphazard manner, suggesting they might have been taken out and placed back in a hurry or for the purpose of the photo.

Question: What type of food is to the left of the baby that is sitting atop the woman?
Label: marshmallow
Prediction: marshmallows
Semantic_type: rel
Structural_type: query
Text: The scene depicts a domestic setting, likely a kitchen, where a woman is seated with a baby. Atop a table, there's a cake adorned with marshmallows and other decorations. The baby seems to be curiously looking at the cake.

Question: Do you see any cats?
Label: yes
Prediction: yes
Semantic_type: obj
Structural_type: verify
Text: The image showcases a Siamese cat with a predominantly cream-colored coat and dark face, sitting next to a pair of worn-out shoes. The cat appears to be curiously inspecting the shoes, and in the background, there's a person wearing a white shirt, sitting on a chair, with a potted plant visible.

Question: Are both the shorts and the black leggings made of cloth?
Label: yes
Prediction: yes
Semantic_type: obj
Structural_type: logical
Text: The image showcases a bowl of soup with a spoon, placed on a plate, next to a stack of toasted bread. The background is dark, possibly a table or countertop.

Figure 2. Examples from **correct** predictions in case that QD image captioning is applied.

Question: Which place is it?
Label: city
Prediction: southwark bridge
Semantic_type: global
Structural_type: query
Text: The image showcases a street sign on a brick building, indicating the name of the street as 'Southwark Bridge Rd.' and its postal code as 'A 300'. The setting appears to be urban, possibly in a historic or older part of a city.

Question: Which place is it?
Label: pasture
Prediction: rural area
Semantic_type: global
Structural_type: query
Text: The image showcases a pastoral scene in a rural area. The primary focus is on a brown cow grazing on a grassy hill. There are other cows in the background, and a wooden structure, possibly a barn or shed, is visible in the distance. The landscape is lush with green grass and trees, suggesting a fertile and well-maintained pasture.

Question: What type of place is pictured?
Label: walkway
Prediction: residential area
Semantic_type: global
Structural_type: query
Text: The image showcases a residential area with a road alongside a water body. There are two children, one on a bicycle and the other on a tricycle, near a stop sign. The trees are adorned with autumn leaves, and there are utility poles and power lines in the background.

Question: Are the words large and black?
Label: no
Prediction: yes
Semantic_type: attr
Structural_type: logical
Text: The image showcases a young individual seated in a large, black-colored chair inside a vehicle. The chair has the words 'TR COMMANDING OFFICER' embroidered on it. The person is wearing a uniform with a badge and is surrounded by various control panels and equipment, indicating that they might be in a commanding or operational position of the vehicle.

Question: Do the tall books look colorful and thick?
Label: no
Prediction: colorful
Semantic_type: attr
Structural_type: logical
Text: The scene in the image showcases a cozy home office setting. A ginger and white cat lounges comfortably on an open laptop. Behind the cat, there's a tall wooden bookshelf filled with various books, some of which are colorful. To the right, there's a bottle of water and some other miscellaneous items on the table. The overall ambiance suggests a relaxed and homely environment.

Question: Are there green snowboards or rackets?
Label: yes
Prediction: rackets
Semantic_type: obj
Structural_type: logical
Text: The image showcases a male tennis player in action on a tennis court. He is about to hit a tennis ball with his racket. The court is green, and there's a chain-link fence in the background.

Question: Are there glasses or women?
Label: yes
Prediction: women
Semantic_type: obj
Structural_type: logical
Text: The image showcases a young baseball player in the midst of a swing, wearing a helmet and a jersey with the number '2'. In the background, there are spectators, including women, watching the game.

Question: Do you see a large mirror or table?
Label: no
Prediction: mirror
Semantic_type: obj
Structural_type: logical
Text: The image showcases a well-lit kitchen with wooden cabinets, a blue countertop, and a dining area. A large mirror is mounted on the wall, reflecting the room and the dining table. The dining table is set with sunflowers and is accompanied by four chairs.

Question: Are there cars or buses?
Label: yes
Prediction: buses
Semantic_type: obj
Structural_type: logical
Text: The image showcases a vibrant and colorful double-decker bus, labeled 'Slowbus' and 'www.slowbus.de', parked on a city street. There are pedestrians walking by, and a car is parked nearby. The backdrop consists of buildings, suggesting an urban setting.

Question: What is common to the door and the room?
Label: color
Prediction: glass door
Semantic_type: attr
Structural_type: compare
Text: The image showcases a spacious and modern living room with a large glass door that leads to an outdoor patio. The room is well-lit, with a predominant white color scheme, complemented by a red sofa and matching chairs. There's a dining area with a glass table and chairs, and a vase with red flowers adds a touch of color to the room.

Question: What do both the soccer ball and the headband have in common?
Label: shape
Prediction: color yellow
Semantic_type: attr
Structural_type: compare
Text: The image showcases a soccer match where a goalkeeper, wearing a yellow jersey and a headband, is in the process of catching or controlling a red soccer ball on the field.

Question: What do the watch and the bicycle have in common?
Label: material
Prediction: outdoors blending
Semantic_type: attr
Structural_type: compare
Text: The scene in the image showcases two individuals outdoors, with one person pouring a liquid into a blender placed on top of yellow crates. There are bicycles in the background, and the person on the right is wearing a watch. The common elements in the image include the outdoor setting, bicycles, and the act of blending or mixing a drink.

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 $\text{OPT}_{2.7b}$, 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

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 1

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 4

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[10] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert. *Zenodo*, 2020. 3

[11] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023. 1

[12] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 2

[13] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019. 2

[14] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 2

[15] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. 1

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2, 3

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1, 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[24] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019. 2

[25] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[26] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[27] OpenAI. Chatgpt. https://openai.com/blog/chatgpt, 2023. 2

[28] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[31] Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700, 2024. 2, 3

[32] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy

Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017. 1, 2

[33] Himanshu Sharma and Anand Singh Jalal. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*, 116:104327, 2021. 1

[34] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. Visual question answering using deep learning: A survey and performance analysis. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*, pages 75–86. Springer, 2021. 1

[35] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. 2

[36] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[37] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 1

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[39] Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[40] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2, 3

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[42] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3081–3089, 2022. 1

[43] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018. 2

[44] Zelin Zhao, Karan Samel, Binghong Chen, et al. Proto: Program-guided transformer for program-guided tasks. *Advances in neural information processing systems*, 34:17021–17036, 2021. 2