

EXPLORING THE CAPABILITIES OF LARGE LANGUAGE MODELS IN VISUAL QUESTION  
ANSWERING: A NEW APPROACH USING QUESTION-DRIVEN IMAGE CAPTIONS AS  
PROMPTS

A THESIS SUBMITTED TO  
THE GRADUATE SCHOOL OF INFORMATICS OF  
THE MIDDLE EAST TECHNICAL UNIVERSITY  
BY

ÖVGÜ ÖZDEMİR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF SCIENCE  
IN  
THE DEPARTMENT OF MODELING AND SIMULATION

SEPTEMBER 2024



**Exploring the Capabilities of Large Language Models in Visual Question Answering: A New Approach Using Question-Driven Image Captions as Prompts**

submitted by **ÖVGÜ ÖZDEMİR** in partial fulfillment of the requirements for the degree of **Master of Science in Modeling and Simulation Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç  
Dean, **Graduate School of Informatics**

\_\_\_\_\_

Assoc. Prof. Dr. Elif Sürer  
Head of Department, **Modeling and Simulation**

\_\_\_\_\_

Assoc. Prof. Dr. Erdem Akagündüz  
Supervisor, **Modeling and Simulation, Middle East Technical University**

\_\_\_\_\_

**Examining Committee Members:**

Prof. Dr. Alptekin Temizel  
Modeling and Simulation, Middle East Technical University

\_\_\_\_\_

Assoc. Prof. Dr. Erdem Akagündüz  
Modeling and Simulation, Middle East Technical University

\_\_\_\_\_

Assoc. Prof. Dr. Hacer Yalım Keleş  
Computer Engineering, Hacettepe University

\_\_\_\_\_

**Date: 05.09.2024**



**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Surname: Övgü Özdemir**

**Signature :**

# ABSTRACT

## EXPLORING THE CAPABILITIES OF LARGE LANGUAGE MODELS IN VISUAL QUESTION ANSWERING: A NEW APPROACH USING QUESTION-DRIVEN IMAGE CAPTIONS AS PROMPTS

Özdemir, Övgü

M.S., Department of Modeling and Simulation

Supervisor: Assoc. Prof. Dr. Erdem Akagündüz

September 2024, 64 pages

Visual question answering (VQA) refers to the artificial intelligence task of providing natural language answers to natural language questions based on a visual input. Due to its requirement for multimodal processing, VQA is considered a challenging problem. It often necessitates understanding the scene depicted in the image, extracting the relationships between objects and their attributes, and performing multi-step reasoning. Over the past few years, numerous deep learning architectures have been proposed for VQA. More recently, pre-trained vision-language models and Multimodal Large Language Models (MLLMs) with billions of parameters have demonstrated superior performance on benchmark tests. Nevertheless, there are still gaps for performance improvement in zero-shot VQA. Zero-shot VQA requires adapting to tasks without input-output guidance, necessitating advanced reasoning abilities. Consequently, recent research has focused on designing prompts that can elicit reasoning capabilities in MLLMs. This thesis proposes a new approach aimed at improving performance in zero-shot VQA by using LLMs and integrating context-aware image captioning as an intermediate step. Evaluations and comparisons were conducted using recent MLLMs, such as CogVLM, GPT-4, and GPT-4o, on the GQA test set, which includes structurally and semantically diverse questions that often require multi-step reasoning. Additionally, the thesis examines the impact of different prompt designs on VQA performance. The findings highlight the potential of using image captions and optimized prompts to enhance VQA performance under zero-shot settings.

Keywords: visual question answering, large language models, image captioning, prompt design, zero-shot VQA

# ÖZ

## BÜYÜK DİL MODELLERİNİN GÖRSEL SORU YANITLAMA YETENEKLERİNİN KEŞFEDİLMESİ: SORU ODAKLI GÖRÜNTÜ ALTYAZILARINI İSTEM OLARAK KULLANAN YENİ BİR YAKLAŞIM

Özdemir, Övgü

Yüksek Lisans, Çokluortam Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. Erdem Akagündüz

Eylül 2024, 64 sayfa

Görsel soru cevaplama (VQA), bir görsele dayanarak doğal dildeki sorulara doğal dilde yanıtların verildiği yapay zeka görevini ifade eder. Çok modlu işlemeyi gerektirmesinden dolayı VQA zorlu bir problemdir. VQA, sıklıkla görseldeki sahneyi anlamayı, nesnelere ve nitelikler arasındaki ilişkiyi çıkarmayı ve çok adımlı muhakeme yapmayı gerektirir. Geçtiğimiz yıllar boyunca, VQA için pek çok farklı derin öğrenme yapısı önerilmiştir. Son yıllarda ise önceden eğitilmiş görsel-dil modelleri ve milyarlarca parametrelili çok modlu büyük dil modelleri (MLLM) değerlendirmelerde üstün performans göstermektedir. Buna rağmen, sıfır-atış VQA'de hala performans iyileştirmek için boşluklar vardır. Sıfır-atış VQA, girdi-çıkış rehberliği olmadan göreve adapte olmayı ve bunun için de gelişmiş muhakeme yeteneği gerektirir. Bu nedenle, son yıllarda araştırmalar, MLLM'ler için muhakemeyi ortaya çıkaracak istem tasarımlarına yoğunlaşmıştır. Bu tez, LLM'leri kullanarak ve bağlama duyarlı görüntü altyazılamayı ara bir adım olarak entegre ederek, sıfır-atış VQA'deki performansını artırmayı amaçlayan yeni bir yaklaşım önermektedir. CogVLM, GPT-4 ve GPT-4o gibi yeni MLLM'ler kullanılarak yapısal ve anlamsal açıdan çeşitli sorular içeren ve genellikle çok adımlı muhakeme gerektiren GQA test seti üzerinde değerlendirme ve karşılaştırma yapılmıştır. Ayrıca, tez, farklı prompt tasarımlarının VQA performansına etkisini incelemektedir. Bulgular, sıfır-atış ayarlarında VQA performansını artırmak için görüntü açıklamalarının ve optimize edilmiş istemlerin kullanım potansiyelini vurgulamaktadır.

Anahtar Kelimeler: görsel soru cevaplama, büyük dil modelleri, görüntü altyazılama, istem tasarımı, sıfır-atış VQA



*To curiosity and hope...*

## ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Erdem Akagündüz, for his invaluable efforts and support. His knowledge and energy have guided and encouraged me throughout my study. I have learned so much from him during this journey, and I am truly grateful.

I also would like to sincerely thank the jury members, Prof. Alptekin Temizel and Assoc. Prof. Hacer Yalım Keleş, for their insightful feedback and contributions.

I would like to extend my heartfelt thanks to Aslı for always lightening my burden and cheering me up during the toughest moments.

Finally, I would like to express my deep gratitude to my parents, Nevzat Özdemir and Nejla Özdemir, and my sister, İstem Özkan, for their constant support and love throughout this journey, just as they have in every journey.

The credits provided by OpenAI through the Researcher Access Program were gratefully acknowledged for the development of this thesis.

# TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZ.....	vi
DEDICATION.....	vii
ACKNOWLEDGMENTS.....	viii
TABLE OF CONTENTS.....	ix
LIST OF TABLES.....	xiii
LIST OF FIGURES.....	xiv
LIST OF ABBREVIATIONS.....	xvi
CHAPTERS	
1 INTRODUCTION.....	1
1.1 Background and Problem Definition.....	1
1.2 Motivation.....	2
1.3 Research Questions.....	2
1.4 Contributions of the Study.....	3
1.5 Organization of the Thesis.....	3
2 LITERATURE REVIEW.....	5
2.1 Visual Question Answering.....	5

2.1.1	Early Approaches .....	6
2.1.2	Transformer-based Approaches .....	6
2.1.3	Neuro-symbolic Approaches .....	7
2.1.4	Graph-based Approaches .....	8
2.1.5	Reasoning in VQA .....	9
2.1.6	Other Related Tasks .....	9
2.1.6.1	Image Captioning .....	10
2.1.6.2	Question Answering .....	10
2.1.7	VQA Datasets .....	10
2.1.7.1	VQA .....	11
2.1.7.2	VQA v2.0 .....	11
2.1.7.3	CLEVR .....	11
2.1.7.4	Visual Genome .....	11
2.1.7.5	OK-VQA .....	11
2.1.7.6	TextVQA .....	12
2.1.7.7	GQA .....	12
2.2	Large Language Models .....	13
2.2.1	Multimodal Large Language Models .....	14
2.2.2	Prompting .....	15
2.2.3	LLM Configuration .....	16
3	METHOD AND IMPLEMENTATION .....	17
3.1	Implementation .....	17
3.1.1	Dataset .....	17

3.1.2	Experimental Setup	19
3.1.3	Evaluation Metric	19
3.2	Initial Analysis: Attention Visualization for Exploring Multimodal Reasoning	20
3.3	Methodology	21
3.3.1	Image Captioning-Integrated VQA Pipeline	21
3.3.1.1	Vision-Language Models for Caption Generation	23
3.3.1.2	Caption Generation Process	24
3.3.1.3	Question Answering	25
3.3.1.4	Baseline VQA Models	26
3.3.1.5	Prompt Details	26
3.4	Results	27
3.4.1	Usefulness of Question-Driven Image Captions	28
3.4.2	Effect of Image Captioning Model	29
3.4.3	Effect of Filtering Image Captions	30
3.4.4	Effect of the Question-Answering Model	30
3.4.5	Predictability of Answers in the Test Set	31
3.4.6	Comparison with the Baseline VQA Models	33
3.4.7	VQA Evaluation using LLMs	33
3.5	Error Analysis	35
3.6	Experiments with Different Prompt Designs	37
3.6.1	Error Analysis with GPT-4o	37
3.6.2	Designing GPT-4o Prompts for VQA	39
3.6.2.1	Adding Object Identification and Hints to the Prompt	39

3.6.2.2	Adding a Vocabulary Definition Instruction to the Prompt .....	41
3.6.3	Results of Prompt Design Experiments .....	42
3.7	Limitations .....	43
3.8	Summary and Discussion .....	44
4	CONCLUSION AND FUTURE WORK .....	47
	REFERENCES .....	49
APPENDICES		
A	ADDITIONAL EXAMPLES OF INCORRECT AND CORRECT PREDICTIONS .....	57
A.1	Predictions Obtained with the QD-IC Integrated VQA Approach .....	57
A.1.1	GPT-4o QD-IC + GPT-4o QA .....	57
A.1.2	CogVLM-C QD-IC + GPT-3.5 QA .....	59
A.2	Predictions Obtained with the standard VQA pipeline .....	61
A.2.1	GPT-4o VQA .....	61
B	ADDITIONAL RESULTS .....	63
B.1	Performance Evaluation Using Cosine Similarity Threshold of 0.8 and Exact Match Criteria .....	63
B.2	Comparison of GPT-4o QD-IC + GPT-4o QA with GPT-4o VQA in Question Categories	64

## LIST OF TABLES

Table 1	Overview of question types according to semantics and structure . . . . .	12
Table 2	Number of samples in <i>test-dev-balanced</i> across question categories . . . . .	18
Table 3	Comparison of IC-integrated and VQA pipelines evaluated using GPT-4o. . . . .	25
Table 4	Comparison of the accuracy values of various IC-integrated VQA pipelines at cosine similarity 0.7, containing the proposed question-driven image captioning (QD-IC) and general-purpose image captioning (GP-IC) on GQA <i>test-dev-balanced</i> . . . . .	28
Table 5	Comparison of the performances of VQA and QD-IC + QA pipelines . . . . .	34
Table 6	Comparison of overall accuracy in IC-integrated VQA pipelines and conventional VQA pipelines for exact matching (EM) and different cosine similarity thresholds. . . . .	34
Table 7	Comparison of IC-integrated and VQA pipelines evaluated using GPT-4o. . . . .	35
Table 8	Comparison of the performance of GPT-4o with different prompts . . . . .	43
Table 9	Comparison of the accuracy values at cosine similarity threshold 0.8 for various IC-integrated VQA pipelines on GQA <i>test-dev-balanced</i> . . . . .	63
Table 10	Comparison of the EM accuracy values for various IC-integrated VQA pipelines on GQA <i>test-dev-balanced</i> . . . . .	63

## LIST OF FIGURES

Figure 1	Examples from GQA dataset. . . . .	13
Figure 2	Proportion of semantically and structurally different question combinations . . . . .	19
Figure 3	Visualization of encoder-decoder multi-head attention weights . . . . .	22
Figure 4	VQA pipeline exploiting general and the proposed question-driven image captioning (QD-IC) as an intermediate step. . . . .	23
Figure 5	Prompt for QD-IC with GPT-4 . . . . .	26
Figure 6	Prompt for QA with GPT-3.5 . . . . .	27
Figure 7	Prompt for QA with GPT-4o . . . . .	27
Figure 8	Prompt for VQA with MLLMs . . . . .	27
Figure 9	Examples of accurate predictions by <i>CogVLM QD-IC + GPT-3.5 QA</i> . . . . .	30
Figure 10	Examples of incorrect predictions of <i>GPT-3.5</i> in the <i>object</i> question category during the QA on <i>CogVLM-C QD</i> image captions. It is observed that <i>GPT-4o</i> correctly answers the same questions. . . . .	31
Figure 11	Examples of <b>correct</b> predictions by <i>GPT-4o QD-IC + GPT-4o QA</i> . . . . .	32
Figure 12	Prompt for VQA Evaluation using LLMs . . . . .	35
Figure 13	Distribution of incorrect and correct predictions of <i>GPT-4o QD-IC + GPT-4o QA</i> across question categories . . . . .	36
Figure 14	Examples of acceptable predictions that do not match the labels . . . . .	37
Figure 15	Examples of predictions containing "unknown" or "unclear" . . . . .	38
Figure 16	Incorrect predictions vs correct predictions of <i>GPT-4o Vision</i> across question combinations . . . . .	39
Figure 17	Examples of predictions containing "I don't know" or "Unknown" . . . . .	40
Figure 18	Examples of acceptable predictions that did not match the labels . . . . .	40
Figure 19	Examples of predictions giving more detailed and correct answers but did not match the labels . . . . .	41



Figure 20	Adding object-attribute identification step and hints to the prompt . . . . .	42
Figure 21	Adding vocabulary definition step to the prompt . . . . .	42
Figure 22	Examples of <b>correct</b> predictions by <i>GPT-4o QD-IC + GPT-4o QA</i> . . . . .	57
Figure 23	Examples of <b>incorrect</b> predictions by <i>GPT-4o QD-IC + GPT-4o QA</i> . . . . .	58
Figure 24	Examples of <b>correct</b> predictions by <i>CogVLM-C QD-IC + GPT-3.5 QA</i> . . . . .	59
Figure 25	Examples of <b>incorrect</b> predictions by <i>CogVLM-C QD-IC + GPT-3.5 QA</i> . . . . .	60
Figure 26	Examples of <b>correct</b> predictions by <i>GPT-4o VQA</i> . . . . .	61
Figure 27	Examples of <b>incorrect</b> predictions by <i>GPT-4o VQA</i> . . . . .	62
Figure 28	Number of correct predictions of <i>GPT-4o QD-IC + GPT-4o QA</i> (blue) vs <i>GPT-4o VQA</i> (yellow) across different structural-semantic question categories . . . . .	64

## LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapping Language-Image Pre-training
CNN	Convolutional Neural Networks
DL	Deep Learning
GRU	Gated Recurrent Unit
GNN	Graph Neural Networks
GPT	Generative Pre-training Transformer
IC	Image Captioning
LLM	Large Language Models
LSTM	Long Short Term Memory
MLP	Multilayer Perceptron
MLLM	Multimodal Large Language Models
NLP	Natural Language Processing
OCR	Optical Character Recognition
RNN	Recurrent Neural Networks
QA	Question Answering
VQA	Visual Question Answering

# CHAPTER 1

## INTRODUCTION

### 1.1 Background and Problem Definition

One of the excellent capabilities of the human brain is multimodal processing, which integrates information from various sensory modalities such as audio, text, image, and video. Even though substantial advancements have been made in Artificial Intelligence (AI), addressing multimodality remains a significant challenge that has drawn attention from academia for the past ten years. Technical difficulties in multimodal processing are mainly associated with the representation, alignment, and integration of multimodal data [1].

One particularly challenging multimodal task is Visual Question Answering (VQA), which lies at the intersection of the fields of Computer Vision (CV) and Natural Language Processing (NLP). The VQA task is defined as providing answers to natural language questions about an image. VQA requires understanding the question, identifying the visual details being asked in the question, and generating a proper answer. VQA often necessitates multi-step reasoning and visual reasoning involving multiple logical deductions based on the elements within the scene and their interrelationships [2]. From this perspective, research on VQA [3, 4, 5, 2, 6] can help understand and reflect the key aspects of human intelligence, specifically merging visual and linguistic information to facilitate reasoning about the environment. VQA systems can also be exploited in various real-life applications, including assisting visually impaired individuals, video surveillance, and educational tools [7].

Recently, Large Language Models (LLMs), developed using transformer architecture and trained on extensive web-scale text corpora, have revolutionized the field of NLP. These models, including GPT-3 and GPT-4 [8], Llama [9], Gemini [10], have demonstrated reasoning capabilities across a wide range of tasks, including language translation, summarization, question answering, and more. Expanding the capabilities of LLMs to connect visual and language modalities has led to the development of Multimodal Large Language Models (MLLMs) [11, 12, 13, 8, 14]. MLLMs indicate high performance in various multimodal tasks, including VQA and image captioning. Although LLMs exhibit strong performance on various tasks, it remains ambiguous whether their decision-making processes truly amount to "reasoning" and to what degree [15]. Hallucination issues are a common problem for both LLMs and MLLMs, impeding the models' capacity for proper reasoning [16]. Therefore, it is still a challenge for MLLMs to execute complex tasks requiring multi-step reasoning without prior examples. This issue is especially evident in zero-shot VQA, where models must answer questions about images without any guiding examples.

LLMs have the ability to perform tasks in few-shot and zero-shot settings, where they are guided by a few examples or only specific task instructions in a process known as "prompting" [17]. Recent research extensively explores Chain-of-Thought (CoT) prompting [18, 17, 19, 20], an approach that enhances reasoning by incorporating intermediate steps producing rationales for the output, applicable to both text-only LLMs and MLLMs. Despite increasing research efforts and promising developments, MLLMs still struggle with zero-shot VQA tasks requiring multi-step reasoning, highlighting the need for new techniques to effectively leverage these models.

This thesis proposes a new approach aimed at improving the zero-shot performance of MLLMs in VQA. To this aim, the study investigates the effectiveness of using context-aware image captions as an intermediate step in the VQA pipeline by feeding them to text-only LLMs as prompts along with the question. The proposed approach also aims to make reasoning in VQA more transparent through the context-aware image descriptions used as rationale. The study conducts a comprehensive evaluation by including various LLMs and MLLMs and using the GQA dataset, which consists of questions involving multi-step reasoning, across different semantic and structural question categories. By leveraging context-aware image captions, the study seeks to bridge the gap in the current capabilities of LLMs, enhancing their performance in complex zero-shot VQA tasks.

## 1.2 Motivation

The motivation for this study stems from the growing importance of developing AI systems capable of multimodal reasoning, a key step towards achieving Artificial General Intelligence (AGI). As AI systems become more integrated into daily life, their ability to understand and reason about multimodal information in a human-like manner becomes crucial.

In this study, the performance of LLMs and MLLMs in zero-shot VQA containing diverse questions requiring multi-step reasoning has been comprehensively evaluated. The purpose of this evaluation is to identify deficiencies in the VQA capabilities of current LLMs and MLLMs and provide an analysis, which can be considered for future improvements.

The approach proposed in this study aims to enhance zero-shot VQA performance by using context-aware image captions as intermediate outputs and providing them as prompts to LLMs. The proposed approach can be useful from several aspects. First, it can be applied to achieve better results in specific question categories within zero-shot VQA. Second, it can be applied in scenarios where the image cannot be provided directly due to privacy concerns. Furthermore, because the LLM uses image captions as rationale, the proposed approach can also be employed in scenarios requiring more interpretable answers.

## 1.3 Research Questions

This thesis will seek answers to the following research questions:

- What is the usefulness of utilizing image captions instead of images within the VQA pipeline, particularly in achieving success in zero-shot VQA?

- Can context-aware image captioning serve as an effective intermediary between images and LLMs?
- Do LLMs exhibit improved performance in the question-answering phase when they rely solely on textual input?
- Can exploiting image captions within a VQA pipeline be advantageous for answering certain types of questions?
- Can prompt designs that encourage MLLMs to engage in multiple thinking enhance performance in VQA?

#### **1.4 Contributions of the Study**

The primary contribution of this study is the proposal of a novel approach termed the question-driven image captioning-integrated VQA pipeline, which aims to establish a connection between images and LLMs for VQA in zero-shot settings.

This study evaluates various image captioning models within the proposed approach to assess their ability to generate informative captions that sufficiently address a range of questions in terms of both semantics and structure.

Additionally, the study examines the performance of various MLLMs on the VQA task, comparing their effectiveness to that of image captioning-integrated VQA pipelines.

Finally, the study explores the effectiveness of various prompt designs that contain preliminary stages within, to enhance the performance of MLLMs in the VQA task.

#### **1.5 Organization of the Thesis**

The thesis is organized into four main chapters:

Chapter 1 introduces the problem definition, outlining the primary challenges in VQA and the objectives of the study. It details the contributions of the study and presents the key research questions guiding the investigation.

Chapter 2 provides an overview of the literature, encompassing various approaches proposed for VQA, including both conventional and recent methods. It also reviews the datasets commonly used in VQA research and presents an overview of LLMs and MLLMs, and different prompting strategies.

Chapter 3 details the implementation step and dataset used in this study. It introduces the proposed question-driven image captioning-integrated VQA approach, followed by experiments that compare the performance of the proposed method with various LLMs and MLLMs. The chapter also examines the contribution of different prompt designs for LLMs in VQA. This chapter also presents evaluation results and includes a discussion of the results.

Chapter 4 highlights the conclusions drawn from experiments with LLMs and MLLMs on the GQA dataset. The chapter highlights future research directions for improving VQA and LLM evaluation.

## CHAPTER 2

### LITERATURE REVIEW

This chapter provides an overview of the Visual Question Answering (VQA) literature, examining various architectures and datasets proposed for this multimodal task. The chapter discusses early methods utilizing various multimodal fusion techniques, as well as more recent approaches, including transformer-based, neuro-symbolic, and graph-based architectures. Furthermore, the chapter explores the latest advancements in Large Language Models (LLMs), including Multimodal Large Language Models (MLLMs) and various prompting strategies.

#### 2.1 Visual Question Answering

Visual Question Answering (VQA) is the task of answering questions posed in natural language based on a visual input. VQA requires handling multimodality and a wide range of skills to address various types of questions [3]. VQA encompasses multiple computer vision sub-tasks such as object detection/recognition, counting, attribute classification, and activity recognition, and measures cognitive abilities such as spatial, common-sense, and knowledge-base reasoning [21].

The challenge of VQA lies in the necessity of establishing a connection between language and visual modalities. It requires understanding the question's requirements and applying reasoning processes by interacting between two different modalities accordingly. This complexity adds challenges to model training, as it requires the model to effectively utilize the interactions between vision and language modalities. Over the past years, different approaches have been proposed to address multimodality in VQA.

A conventional VQA pipeline usually consists of three main stages: image featurization, question featurization, and multimodal fusion [22].

In the image featurization step, the input image is processed to extract visual representations, typically by passing the image through a pre-trained model based on a Convolutional Neural Network (CNN) structure such as ResNet [23] or VGG [24].

In question featurization, the input question is processed to capture its semantic representation. This involves encoding the question into a vector representation that the model can process. Models based on Recurrent Neural Networks (RNN) structure such as Long Short-Term Memory (LSTM) [25] and Gated Recurrent Unit (GRU) [26] are typically used to convert the question into a fixed-length vector representation.

During the fusion step, visual and textual features are transformed into a shared feature space and combined using various techniques such as concatenation, element-wise summation, element-wise multiplication, compact bilinear pooling, and cross-modal interactions.

### **2.1.1 Early Approaches**

In early studies, CNN-based models were predominantly used for image featurization, while RNN-based models were commonly employed for question featurization, with the main distinction between approaches occurring at the fusion stage.

The fusion of visual and linguistic data is the key element of VQA. This requires uncovering the complex and high-level interactions between the textual semantics in the question and the visual concepts depicted in the image.

Gao et al. [27] propose a VQA model extracting visual features with CNN and language features with LSTM, and fusing them with element-wise addition. Joint embeddings are further fed into a classifier to predict an answer.

Kim et al. [28] introduce Multimodal Residual Networks (MRN) that learn joint representation from vision and language representations. The fusion of two modalities is carried out using element-wise multiplication.

Ben-younes et al. [29] introduce MUTAN using multimodal Tucker fusion to handle the fusion of textual and visual modalities. MUTAN learns joint representations of images and questions in a shared embedding space. The Tucker decomposition is utilized to factorize the acquired multimodal representations into core tensors and mode-specific factor matrices. MUTAN employs attention processes to dynamically prioritize relevant sections of the image and questions while merging them.

Feature-wise Linear Modulation (FiLM) [30] proposes a layer for conditioning the visual features based on the input question by applying affine transformations. This process modulates the visual feature maps using parameters derived from the question, allowing the model to focus on relevant aspects of the image associated with the question.

Santoro et al. [31] propose Relation Networks (RN), which structure the neural network's functionality by considering the pairwise relationships among objects. In this approach, images are processed with a CNN to obtain a set of objects for the RN. After calculating the relationships between all pairs of objects, the Relation Network aggregates these features to create a global representation that encodes relational information to answer the question.

### **2.1.2 Transformer-based Approaches**

Transformer architecture was first introduced in the paper "Attention is All You Need" [32] in 2017 by Vaswani et al. Transformers demonstrate remarkable success in a variety of tasks in NLP, such as text summarization, sentiment analysis, and question-answering, and become the foundation of recent LLMs.



The transformer architecture comprises an encoder and a decoder which contains sub-layers with multi-head self-attention and feed-forward neural networks. Self-attention layer, the key element of transformers, analyzes relationships within input sequences and calculates the weights that represent the importance of each word relative to all other words in the sequence. Self-attention allows the model to capture long-term dependencies and relationships between different segments of the input. Multi-headed self-attention enables the learning of different sets of self-attention weights in parallel, with each set focusing on various aspects of the language.

The large-scale pre-training concept is essential in the success of transformer-based models. This involves training transformer models on massive corpora using self-supervised learning techniques. By pre-training on large-scale data, transformer models can learn rich and generalized representations of the input data, which can then be fine-tuned on relatively small amounts of labeled data for specific downstream tasks.

Transformers are initially used in the NLP domain, then extended to vision [33] and multimodal tasks [34, 35, 36]. Cross-modal interactions are essential for tackling multimodality in transformers. These interactions usually involve early summation, early concatenation, hierarchical attention, and cross-attention techniques [37]. Multimodal pre-training is also extensively researched with the emergence of large-scale multimodal datasets. Multimodal pre-training typically involves training with techniques such as image-text matching and contrastive learning. Transformers with large-scale multimodal pre-training surpass the performance of previous architectures in various downstream tasks.

LXMERT [35] is a large-scale transformer model composed of an object-relationship encoder, a language encoder, and a cross-modality encoder that integrates textual and visual semantics. The cross-modality encoder includes both self-attention and cross-attention layers. It processes the outputs from the language encoder and the object-relationship encoder. Cross-attention allows the model to focus on relevant parts of the image while processing the question and vice versa. Similar to self-attention, cross-attention computes attention scores between pairs of tokens from different sequences to create joint cross-modality representations.

OSCAR [36] proposes a Visual-Linguistic Pretraining (VLP) method that enhances cross-modal representation learning by utilizing object tags detected in images as anchor points. The training samples are constructed as triples, each consisting of a word sequence, a set of object tags, and image region features. OSCAR uses masked token loss and contrastive loss as pre-training objectives.

MDETR [38] builds upon the principles of the object detection model DETR [39], enhancing it with the capability to process free-form text inputs in addition to the input image. Similarly to DETR, image features are extracted with a CNN backbone, and text features are encoded with a language transformer RoBERTa [40]. Image and text feature vectors are concatenated and given to the transformer encoder, which is followed by the transformer decoder taking object queries and question-answering specific queries attended to encoder output.

### **2.1.3 Neuro-symbolic Approaches**

Neuro-symbolic AI introduces a hybrid paradigm that seeks to combine the power of neural networks with the structured, rule-based reasoning of symbolic systems. By integrating neural networks with

symbolic reasoning, which can execute precise logical operations, neuro-symbolic AI holds the potential to create systems that are both highly flexible and interpretable.

In VQA research, neuro-symbolic models have garnered significant attention because their compositional structure allows for transparency and interpretability. These models typically include a question parser to generate programs from questions, a scene parser to identify relevant objects and attributes, and a program executor to apply these programs to scene representations.

Neural Module Networks [41] proposes a dynamic structure built by decomposing questions into linguistic sub-structures with a parser and combining necessary computational units, *e.g.*, classification, and attention, that represent the associated sub-structures. Similarly, End-to-end Module Networks [42] proposes an end-to-end approach that learns both parsing the questions and a layout policy with RNNs for dynamically constituting computational units.

NS-VQA [43] employs neural networks for scene parsing and decomposing questions into program instructions. It introduces a symbolic module that executes these program instructions on the scene representation. NS-VQA constitutes scene representations using Mask-RCNN to segment objects in the input image, and a CNN-based model to classify the object attributes with full supervision.

An extended work NS-CL [44] constructs scene representations without fully annotated supervision, unlike NS-VQA. Question parser training, a significant step of the reasoning process, includes the steps of supervised learning of mapping questions to program instructions in a small data set, and then reinforcement learning by pairing the resulting model with a deterministic program executor.

Neuro-symbolic approaches can also integrate with other neural architectures. A hybrid approach, ProTo [45], introduces program-guided transformers that leverage both the semantic and structural information from programs parsed from questions by a sequence-to-sequence model.

A recent approach VisProg [46] uses a language model with in-context examples to generate programs. Programs are then executed using pre-defined modules that perform various functions, including image cropping, object localization, and segmentation. VisProg stands out for its interpretability and adaptability to various other tasks.

While neuro-symbolic models provide more transparent reasoning mechanisms in VQA systems, they frequently require manual design of program execution steps, which limits their scalability and makes it difficult to develop generalized solutions for diverse tasks.

#### **2.1.4 Graph-based Approaches**

In recent years, Graph Neural Networks (GNNs) have garnered significant attention from the community due to their ability to represent interconnections and provide interpretability. GNNs enable the construction of relationships in a scene based on the given question. Several related works employ GNNs for visual question answering.

GraphVQA [6] proposes a language-guided graph neural network for constructing reasoning in VQA. GNNs follow the message-passing scheme, updating each node’s feature using its local neighborhood. The main idea of GraphVQA is to convert a natural language question into several iterations of message passing between graph nodes. Each message passing iteration is performed by a layer of a GNN.

An alternative graph-based approach for VQA is proposed in [47]. The method integrates a graph learner module that generates a question-specific graph representation of the input image with graph convolutions to capture question-specific interactions.

Language-Conditioned Graph Networks (LCGN) [48] is targeted to support relational reasoning in VQA and Referring Expression Comprehension (REF). The model builds contextualized object representations through iterative message passing conditioned on textual input. Each object node updates its representation based on relationships to support relational reasoning.

### 2.1.5 Reasoning in VQA

Reasoning is defined as the process of deriving conclusions or inferences from information [49]. Visual reasoning (VR) refers to the capability to create a comprehensive depiction of a visual scene and carry out multi-step logical deductions concerning the elements within the scene and their interrelationships [2]. As one of the tasks involving visual reasoning, VQA often requires the ability to infer relationships between objects in a scene, recognize relational patterns, and make logical deductions considering both visual and textual inputs. For instance, to correctly answer a question like "What color is the car next to the tree?", the model needs to understand the spatial relationships between objects in the image, identify the car and the tree, and deduce the color of the car through visual analysis.

As mentioned in the previous sections, numerous studies have been proposed to enhance the performance and reasoning capability of VQA models, such as graph neural networks or neuro-symbolic models. Several other studies have examined various aspects of reasoning in VQA, focusing on factors such as neural architecture and input uncertainties that may influence reasoning ability. Xu et al. [50] examine what reasoning tasks a neural network can learn well. The primary concept of the study is based on the observation that reasoning processes are similar to algorithms. Thus, their study investigates how well a reasoning algorithm aligns with the network's computation graph. The authors introduce an algorithmic alignment framework that formalizes the interaction between neural network structures and reasoning processes such as summary statistics, relational argmax, and dynamic programming. Their results highlight that GNNs succeed in various reasoning tasks because their underlying reasoning processes resemble dynamic programming [50].

Kervadec et al. [51] investigate whether uncertainty in vision is a key obstacle to effectively learning reasoning in vision and language problems. They define reasoning as the algebraic manipulation of words and visual objects to predict answers to new questions and argue if reasoning becomes challenging when the visual input is noisy and uncertain, as opposed to learning from clear information about a scene. To test this hypothesis, they trained an oracle model with perfect visual information, providing the model with ground-truth object annotations from the GQA dataset [3]. Their observation indicates that the model with perfect sight has more diverse attention modes, resembling a reasoning process by predicting answers from visual information rather than relying on biases.[51].

### 2.1.6 Other Related Tasks

In this section, a brief overview of two tasks related to VQA is presented: image captioning and question answering. These tasks are addressed in the application part of this thesis as modules within the proposed VQA pipeline.

### **2.1.6.1 Image Captioning**

Image captioning (IC) is the task of generating captions describing the visual content in natural language. Similar to VQA, IC requires integrating visual and language understanding. Early approaches explore retrieval-based methods that usually find the closest matching image from an extensive database of images accompanied by descriptions to produce a caption [52]. Deep learning-based IC relies on attention and encoder-decoder mechanisms [53, 54]. A typical architecture consists of a CNN-based encoder to map visual representations in a shared embedding space and an RNN-based decoder to produce the final caption. Moreover, transformers revolutionizing NLP and vision, are also employed successfully in image captioning by enabling end-to-end training [55, 56].

Alternatively, scene graphs are extracted using different methods and utilized in image captioning [57]. Typically, a scene graph consists of objects, attributes, and the relationships between them. A scene graph serves as an intermediary to bridge the gap between pure-vision and vision-language tasks [55]. Graph Convolutional Networks and Long Short-Term Memory (GCN-LSTM) [58] are combined to model visual relationships and produce associated captions based on these relationships. Yang et al. [59] integrate scene graphs into the encoder-decoder structure as a way of incorporating inductive bias.

Recently, MLLMs [60, 46, 12, 8, 10, 13] have significantly advanced the field of image captioning, similar to their contributions in various other vision-language tasks. MLLMs are discussed in detail in Section 2.2.1

### **2.1.6.2 Question Answering**

Question-answering (QA) is the task of generating contextually appropriate responses to questions in natural language. A traditional QA system typically consists of the steps of document retrieving and reading [61]. The document retrieval process aims to identify sources or documents relevant to the question, while the reading step focuses on extracting the appropriate answer, based on the related sources.

QA research has gained significant attention in recent years, resulting in the development of numerous deep learning-based systems. The Stanford Question Answering Dataset (SQuAD) [62] is extensively utilized in deep learning-based QA research. Recent advancements in QA models cover transformer architectures and large-scale pre-training on diverse datasets [63, 40, 64, 65]. Notably, LLMs such as OpenAI’s ChatGPT [8] and Google’s Gemini [10] have demonstrated substantial improvements in conversational question answering by effectively interacting with humans.

### **2.1.7 VQA Datasets**

This section gives an overview of extensively researched VQA datasets.

### **2.1.7.1 VQA**

Antol et al. [4] propose the VQA dataset and present the concept of free-form and open-ended VQA. The dataset contains about 204K images from the MS-COCO dataset [66], and 760K questions. Subsequent studies have shown that the VQA dataset contains inherent biases where questions can often be answered correctly without looking at the image, based purely on language patterns in the training data.

### **2.1.7.2 VQA v2.0**

VQA v2.0 [67] aims to reduce biases from the previous VQA dataset [4] and balance question-answer pairs. This improvement aims to reduce the chances of models relying on language biases rather than understanding the visual content. The VQA v2.0 dataset has roughly 1.1M image-question pairs with 13M corresponding responses [67].

### **2.1.7.3 CLEVR**

CLEVR (Compositional Language and Elementary Visual Reasoning) [68] dataset is designed to evaluate visual reasoning abilities. It contains 100K synthetic images with various shapes, colors, numbers, and materials of objects and around 853K compositional questions being generated programmatically. These questions are designed to measure skills for counting, comparing, and querying object properties and relationships. To aid visual reasoning research, the dataset also includes scene graph annotations and functional representations. However, due to its synthetic origin and limited variety, characterized by a small number of object categories and qualities, the dataset is stated as susceptible to the memorizing of all possible combinations [3].

### **2.1.7.4 Visual Genome**

With an average of 35 objects, 26 attributes, and 21 pairwise relationships between objects, the Visual Genome dataset [69] comprises more than 108K images. The dataset includes scene graphs of real-world images where objects, attributes, and their relationships are defined. Visual Genome encompasses a wide range of visual tasks from perception to reasoning with 1.7M questions and answers [69].

### **2.1.7.5 OK-VQA**

OK-VQA [70] is proposed to address knowledge-based VQA, which requires using external knowledge beyond the visual content to answer a question. The OK-VQA dataset has approximately 14K questions across several subject domains, including science and technology, history, and sports [44]. Unlike other VQA challenges, OK-VQA necessitates the use of a VQA model to determine the external knowledge required to answer the question and then apply this information to generate an answer.

### 2.1.7.6 TextVQA

TextVQA [71] aims to address the challenge of answering questions that require reading and understanding text embedded in images. The challenge arises in scenes where text is an important part of the visual content, such as street signs or product labels. TextVQA comprises approximately 45K questions across 28K images that necessitate text-based reasoning for generating answers.

### 2.1.7.7 GQA

GQA [3] comprises 113K real-world images and 22M questions being generated by a question engine from scene graph structures in Visual Genome [69] dataset. The dataset evaluates performance on a range of reasoning skills, including object and attribute recognition, tracking transitive relations, spatial reasoning, logical inference, and comparisons [3]. Each image in the dataset is accompanied by a scene graph, and each question is linked to a functional program that specifies the reasoning steps, based on the scene graphs, to generate an answer.

The GQA dataset features a diverse set of questions categorized primarily by their structure and semantics. The structural type of a functional program is determined by its final operation, which can include categories such as *verify*, *query*, *choose*, *logical*, and *compare*. The semantic type defines the main subject of the question and includes categories like object, attribute, category, relation, and global. Descriptions of these question categories, along with examples, are provided in Table 1. Each question is associated with one structural and one semantic question type. Examples from the dataset are shown in Figure 1.

Questions in the GQA dataset often demand multiple reasoning steps. As detailed in the dataset’s accompanying paper, 6% of the questions require one step of reasoning, 42% require two steps, 44% require three steps, and 8% necessitate more than four steps.

Table 1: Overview of question types according to semantics and structure

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

GQA is notable among the VQA datasets due to its focus on evaluating compositional and grounded reasoning skills, as well as its comparatively varied question-answer set. Additionally, the questions in GQA are challenging because they not only pertain to the main focus of the images but also encompass



Figure 1: Examples from GQA dataset.

the details within the visual content. Therefore, the GQA dataset has been selected as a benchmark dataset in the evaluation part of this study.

## 2.2 Large Language Models

With the development of transformers, Natural Language Processing (NLP) has reached an important milestone. Pre-trained language models, *e.g.*, Bidirectional Encoder Representations from Transformers (BERT) [63], trained on large-scale corpora of text, and have led to remarkable progress in many downstream tasks such as summarization, question-answering, and translation. Scaling the model size and data in pre-trained language models has resulted in the development of Large Language Models (LLMs) with improved capacity.

LLMs are usually pre-trained to predict the next token or fill in masked words, depending on their specific objectives. The pre-training is conducted using self-supervised learning, aiming for the LLM to understand the patterns and structures of the language. Recent LLMs are further fine-tuned on a smaller instruction dataset to enhance their ability to follow instructions, which is called *instruction tuning*. Unlike pre-training, instruction tuning is a supervised learning process in which labeled data, *i.e.*, prompt-completion pairs, are required.

A significant advancement for LLMs was the release of GPT-3 [65] by OpenAI in 2020. *In-context learning* paradigm was first utilized in the GPT-3 paper, which refers to adapting to a target task with a given prompt in the natural language during inference. In-context learning allows the LLM to produce the target output without the need for additional training or gradient updates. Pre-training involves predicting the next text sequence based on context, while in-context learning predicts the correct task solution, formatted as a text sequence similarly, using the task description and examples provided [72].

An enhanced version of the GPT-3, InstructGPT [73], was released in 2022, employing *reinforcement learning from human feedback (RLHF)* as an additional fine-tuning process to prevent undesired outputs such as toxicity and falsehoods. Further advancements led to GPT-3.5 and ChatGPT which is a conversation model based on GPT-3.5, that have a greater capacity than GPT-3. ChatGPT has the ability to interact with humans, demonstrating comprehensive knowledge, mathematical reasoning skills, and contextual understanding in multi-turn conversations [72]. GPT-4 [8] was introduced in 2023, which outperforms previous GPT models and enables multi-modal input processing. In May 2024, the most recent model of the GPT family, GPT-4o [74], was released. GPT-4o accepts and produces

multi-modal data such as text, image, audio, and video. OpenAI reported that GPT-4o outperforms GPT-4 in text, reasoning, and coding intelligence benchmarks [74].

Many other LLMs with billions of parameters, such as LLaMA [9], BLOOM [75], PaLM [76], and CogVLM [12] employ various design approaches and demonstrate remarkable performance in benchmarks. Training LLMs with billions of parameters requires substantial computational resources. Therefore, APIs offer a practical solution, enabling developers to access these models without extensive infrastructure. OpenAI APIs for GPT-3.5, GPT-4, and GPT-4o models have been utilized in the practical part of this thesis study.

Although LLMs demonstrate outstanding performance in a diverse set of tasks, hallucinations remain a significant limitation and an ongoing research challenge. Hallucination refers to content that seems fluent and natural but is nonsensical or misaligned with the source material [57]. Several factors can lead to hallucinations such as source-reference divergence, imperfect representation learning, and parametric knowledge bias [57]. Mitigating hallucinations is an open research problem with efforts focused on improving model training, evaluation, and prompt engineering techniques.

### 2.2.1 Multimodal Large Language Models

The human brain can incorporate various inputs, including visual, auditory, and tactile signals. LLMs were originally developed to comprehend human language and perform language-related tasks such as translation, summarization, and text generation. Recently, there has been a significant effort in research to broaden the capabilities of LLMs by integrating additional modalities, including audio and visual inputs. This integration aims to create more versatile models capable of processing, reasoning, and producing multimodal information, hence advancing the progress towards Artificial General Intelligence (AGI).

LLMs having the capability to process and produce multimodal inputs are often referred to as *Multimodal LLMs (MLLMs)* or *Large Multimodal Models (LMMs)*. Numerous neural architectures focus on vision-language integration [35, 36]. Multimodal LLMs are distinguished by their vast scale, often encompassing billions of parameters, and the use of training techniques such as instruction tuning [77]. GPT-4V [8], LLaVa [14], Gemini [10], CogVLM [12] are some of the state-of-the-art MLLMs showcasing advanced capabilities in multimodal processing.

A standard MLLM is composed of three primary elements: a pre-trained modality encoder, a pre-trained LLM, and a modality interface [77]. A modality encoder is usually an encoder pre-trained on large-scale image-text pairs and used for natural language to identify visual concepts [64]. It converts visual and language modalities into a unified representation. A modality interface is responsible for converting information into a format that the LLM can effectively process, usually achieved by token-level and feature-level fusion [77]. Token-level fusion converts features produced by the encoder into tokens. In feature-level fusion, additional modules such as cross-attention layers are added to facilitate deep interaction between text and visual features.

MLLM goes through three training stages: pre-training, instruction tuning, and alignment tuning. A typical pre-training for MLLMs involves keeping the pre-trained modules, i.e., visual encoders and LLMs, fixed while training a learnable modality interface. Instruction tuning focuses on enhancing the model's ability to comprehend user instructions. The data for instruction tuning contains instructions



usually natural language input defining the target task, image-text pairs, and corresponding output for the instruction. Alignment tuning is employed to meet human expectations, through techniques such as *reinforcement learning with human feedback (RLHF)*.

### 2.2.2 Prompting

The input text in natural language fed into an LLM is called a *prompt*. A prompt is given to the LLM at inference time and typically includes instruction, input data, and output format for the associated task. The process of iteratively developing and refining a prompt to achieve the desired output is called *prompt engineering*.

General prompting strategies in OpenAI documentation<sup>1</sup> are summarized as follows:

- Specifying the task and output format clearly
- Providing reference text to prevent fake answers
- Decomposing complicated tasks into simpler tasks
- Creating a *chain-of-thought* or allowing inner monologue to give time to LLM for reasoning

LLMs can rapidly adapt to various tasks at inference time. *Zero-shot prompting* refers to directly asking the LLM to perform a task without providing any examples of input-output pairs at inference time. This approach is required for the model's ability to generalize to new, unseen tasks based only on the instructions provided. *One-shot prompting* provides the model with just one input-output pair as an example, similar to how humans adapt to new tasks. *Few-shot prompting* includes giving a few examples of input-output pairs related to the task to guide the model during inference.

Various prompting strategies have been proposed for enhancing the reasoning capabilities of LLMs during inference. *Chain-of-thought (CoT) prompting* [18] is a technique that decomposes complex tasks into several intermediate reasoning steps, inspired by the human thought process. CoT is exploited as a few-shot prompting technique by providing input, output, and corresponding chain-of-thought processes as exemplars. CoT is shown to be effective in enhancing LLM's performance in tasks including arithmetic reasoning, commonsense reasoning, and symbolic reasoning [18]. Zero-shot CoT [17] proposes an alternative technique to trigger CoT in zero-shot settings by simply adding "Let's think step by step" in the prompt.

Chain-of-thought for language-only LLMs has been further extended with multimodal approaches. Multimodal CoT (MM-CoT) [19] proposes a two-stage framework to incorporate both visual and language input with chain-of-thought prompting. The proposed framework consists of rationale generation and answer inference steps. In the first stage, the model receives both language and vision inputs to produce rationales. In the second stage, these rationales are combined with the original language input. The model then uses this updated language input, along with the original vision input, to produce the answer. Duty-distinct chain-of-thought (DD-CoT) [20] proposes a multimodal CoT prompting approach. DD-CoT first breaks the question into sub-questions and then prompts the LLM to assess

---

<sup>1</sup> <https://platform.openai.com/docs/guides/prompt-engineering>

whether each sub-question can be answered without visual information. Sub-questions are then answered by a VQA model to guide overall image understanding, and supplied with their corresponding sub-answers as additional information to LLMs to engage in joint reasoning.

### 2.2.3 LLM Configuration

Besides the prompt, configuration parameters also play an important role for LLMs to generate the desirable output at inference time. LLMs predict a probability distribution over the sequences of tokens. The functions of these parameters are summarized in the following sentences.

*Sampling method* defines the selection process of tokens over the predicted probability distribution. Widely used sampling methods are *greedy sampling* and *random sampling*. While *greedy sampling* selects the most probable token, *random sampling* randomly selects the next token according to its probability distribution.

*Temperature ( $T$ )* controls the level of randomness in the generation process. Increasing the temperature adds more randomness to the output, while  $T=0$  deterministically selects the token with the highest probability.

*Top- $k$*  limits the vocabulary size where the  $k$  defines the number of tokens with the highest probability.

*Top- $p$* , also called nucleus sampling, specifies a cumulative probability threshold that tokens whose cumulative probabilities fall below are considered. Similar to *top- $k$*  and *temperature*, it impacts the output variety.

*Max length* defines the maximum number of tokens to be generated by the LLM. A similar parameter to control the output size is *stop sequences* which defines the points at which the LLM will halt the generation process.

*Frequency penalty* is a parameter that reduces the likelihood of repeated tokens in the generated text by penalizing tokens based on their frequency of occurrence.

*Presence penalty*, similarly to *frequency penalty*, discourages repetition by penalizing tokens based on their occurrence, regardless of how frequently they appear.

## CHAPTER 3

### METHOD AND IMPLEMENTATION

In this chapter, the infrastructure of the experiments conducted in the study, the proposed methodology, details of the experiments, and the results will be explained. The chapter includes the following sections. In Section 3.1, the dataset, software and hardware setup, and evaluation metrics used in the experiments will be introduced. Section 3.2 will present attention visualizations conducted at the beginning of the study to understand reasoning in pre-trained vision-language transformers. Section 3.3 will describe the proposed image captioning-integrated VQA pipeline, detailing each stage of the process. Section 3.4 will examine the experimental results under different headings. In Section 3.5, the findings of the error analysis will be shared. Section 3.6 will assess the impact of different prompt designs on VQA performance. Section 3.7 will address the limitations of the study. Finally, Section 3.8 will summarize and discuss the results.

#### 3.1 Implementation

##### 3.1.1 Dataset

Experiments were carried out on the GQA dataset [3], specifically using the balanced version of the *test-dev* subset, i.e., *test-dev-balanced*, generated to eliminate question-conditional biases. The subset includes 12,578 questions and 398 images, where each image in the dataset is linked to multiple questions.

GQA includes various types of questions in terms of semantics and structure. Each question is associated with at least one structural and one semantic category. The number of samples in each question type existing in the *test-dev-balanced* is given in Table 2. An overview of the GQA dataset is given in Section 2.1.7.7, along with the definitions of the question types and related examples.

A detailed chart indicating the proportion of semantically and structurally different question combinations in the *test-dev-balanced* subset is given in Figure 2.

Some findings that can be drawn from the bar chart are summarized below.

- The most frequent *structure-semantic* combination in the test set is *query-relation* with 3,887 instances. The *relation* category often refers to the spatial relationship between two objects. An example of a question from the *query-relation* category is "What kind of meat is above the fork?".

Table 2: Number of samples in *test-dev-balanced* across question categories

Category	Question Types	No. Samples
Structural	verify	2252
	query	6805
	choose	1128
	logical	1803
	compare	589
Semantic	object	778
	attribute	5185
	category	1149
	relation	5308
	global	157

- *Attribute* semantic category is combined with all structural categories, mostly with *query* and *logical* categories. *Attribute* questions often inquire about the color, size, or action of an object in the scene. An example question from *query-attribute* category is "What is the material of the chair?", while an example question from the *logical-attribute* is "Does the calf have brown color and large size?".
- Another semantic category which the *query* structural category frequently occurs with is the *category* class. *Query-category* questions typically ask which object possesses a certain attribute, e.g., "What type of clothing is pink?".
- Questions in the *verify* class are mostly associated with the *relation* and *attribute* semantic categories. *Verify-relation* questions often inquire about the existence of an object, focusing on its location relative to another object, e.g., "Is there a blender to the right of the yellow drink?". *Verify-attribute* questions typically ask about the attributes of an object in relation to another object, such as "Does the device under the picture frame look black?".
- Examples in the *global* semantic class are rare, with most occurring alongside the *query* structural category. These questions pertain to the entire scene, such as "Which place is it?" or "How is the weather?".
- *Logical-attribute* and *logical-object* categories show notable counts, while other combinations for the *logical* semantic class are missing. *Logical-attribute* are typical "yes/no" questions asking whether an object related to another object possesses both attributes, such as "Are the boxes to the right of the man full and square?". *Logical-object* questions generally pertain to the existence of objects with a specific attribute in the image, such as "Are there any small refrigerators or microwaves in the image?".
- *Compare* structural category only occurs with the *attribute* semantic category in the dataset. These are typically "yes/no" questions about whether two different objects in the scene share the same attribute, such as size, color, or material. For instance, "Are the sidewalk and the traffic sign made of the same material?".

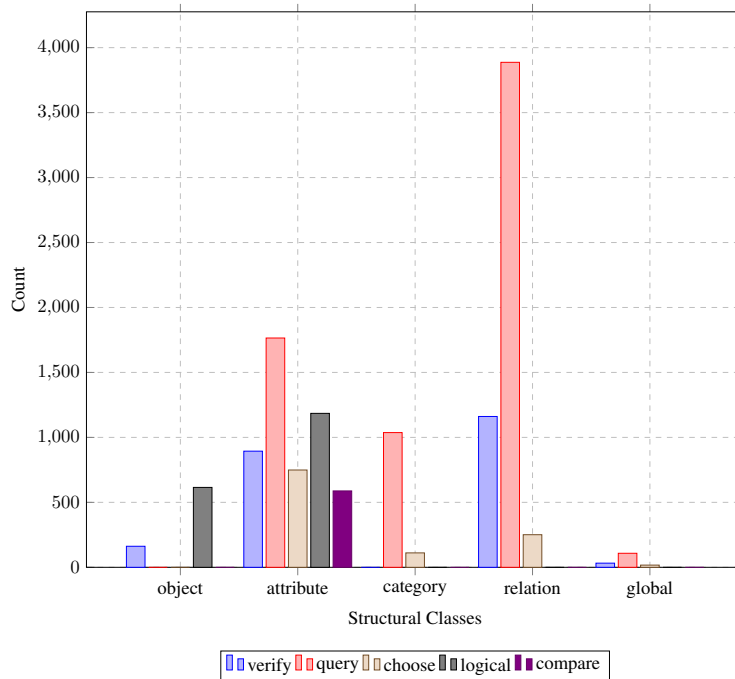


Figure 2: Proportion of semantically and structurally different question combinations

### 3.1.2 Experimental Setup

Python programming language was employed in this project. For LLM/MLLM inference, OpenAI’s APIs were utilized, specifically calling the GPT-3.5, GPT-4, and GPT-4o models with the chat completion role. Pre-trained vision-language models were sourced from HuggingFace, and executed for inference on a computer equipped with an NVIDIA A100 GPU having 16 GB of VRAM capacity. For quantization steps to align the models with our computational resources, the Bitsandbytes library<sup>1</sup> was used.

### 3.1.3 Evaluation Metric

*Accuracy* was considered as the evaluation metric throughout the experiments. *Accuracy* is defined as the ratio of correct predictions to the total number of predictions. Since the answers are produced by LLMs with open-ended generation, it is likely that predictions do not match the ground truths even though they can be accepted as correct. Therefore, cosine similarity was employed to word vectors to evaluate the alignment between the LLM’s predictions and the ground truth answers. Word vectors were generated using the spaCy library with the "en\_core\_web\_md" model. Cosine similarity between word pairs was calculated by normalizing the dot product of their vectors. A threshold of 0.70 was set, with scores meeting or exceeding this threshold indicating semantic similarity. For example, the prediction "couch" is accepted as correct for the ground-truth "sofa", as the semantic similarity between the two words exceeds 0.7. It was observed that at lower thresholds, many predictions being contextually similar but not semantically identical to the ground truth were also accepted as correct.

<sup>1</sup> <https://github.com/bitsandbytes-foundation/bitsandbytes>

For instance, the similarity between "blue" and "brown" is measured as 0.67. The evaluation was also conducted using higher cosine similarity thresholds, such as 0.8 and 0.9, as well as for exact matching (EM).

An alternative answer set (AAS) generated by Luo et al. [78] for GQA *test-dev-balanced*, which comprises ground-truth answers and their semantically close alternatives, was also utilized for evaluation. The generation of the AAS involves leveraging WordNet [79] and ConceptNet [80] knowledge bases, as well as BERT and counter-fitting word embeddings for each ground-truth answer [78]. The Union-AAS metric, which combines alternatives produced using these four different methods, was also included as an evaluation metric.

### 3.2 Initial Analysis: Attention Visualization for Exploring Multimodal Reasoning

As an essential element of transformer architecture, attention mechanism enables the model to weigh importance of different segments of the input elements for a given task. The visualization of attention weights is useful for providing insights into the decision-making process of complex deep learning models, often referred to as black-boxes. In VQA, this visualization reveals which parts of an image the model prioritizes when generating answers to the corresponding questions. Just as humans actively observe and reason in a step-by-step manner to accomplish a visual task, attention and reasoning are closely linked mechanisms that shape the decision-making process [81]. The model's focus on areas that the human eye would concentrate on is an expected outcome when applying a reasoning process similar to that of humans.

This section will examine how transformers handle multimodality through the attention visualization of a pre-trained transformer model, MDETR [38]. The authors state that MDETR is pre-trained on 1.3M text-image pairs sourced from several multimodal datasets with alignments between textual phrases and visual objects [38]. After pre-training, the model is fine-tuned for several downstream tasks, including VQA. The fine-tuned models are available on the GitHub repository<sup>2</sup> of the study.

MDETR employs a convolutional backbone to capture visual features and a language model to derive text features. These features, derived from both modalities, are combined in a shared embedding space and fed into a transformer encoder-decoder. MDETR's encoder processes concatenated image and text features to learn cross-modal interactions. The decoder takes object queries and QA-specific queries, encoded as learned embeddings, to generate predictions for the bounding boxes of the objects referenced in the question, as well as to determine the question type and the appropriate response for each question category. 100 object queries, 1 question-type query, and 5 answer-specific queries are fed into the decoder.

The authors fine-tuned MDETR using the EfficientNet [82] and the ResNet-101 [23] backbones on the GQA dataset, achieving an accuracy of 62.45% and 61.99% on the test-dev subset, respectively. In this analysis, the MDETR model, fine-tuned on the GQA dataset with an EfficientNet-B5 backbone, has been utilized through attention visualization. In the analysis, the attention weights from the decoder's final layer averaged over all heads have been extracted. Totally, six attention maps specifying the question-type, and answers for *object*, *relation*, *attribute*, *category*, and *global* question types have been visualized. Examples from attention maps are given in Figure 3.

---

<sup>2</sup> <https://github.com/ashkamath/mdetr>

The maps on the right of the figure display the attention weights corresponding to predictions for the following categories: question type (q\_type), object (obj), relation (rel), attribute (attr), category (cat) and global. In this analysis, the attention weights generated while producing the answer for the relevant question type are considered based on the predicted question type. For example, in Question 1 in Figure 3, since the question type is predicted as "attr", the attention weights in the last layer of the decoder generated for the answer corresponding to the "attr" question type are taken into account in the analysis.

The first 100 attention maps were manually analyzed, revealing the following results: In 50 of the questions, the model focused on the correct parts of the image and predicted the correct answer. For example, for Question 1 in Figure 3 about the characteristics of the tomato, the model accurately targeted the relevant area of the image. In 18 of the questions, the model predicted the correct answer despite focusing on irrelevant parts of the image. For example, for Question 3 in Figure 3, despite focusing on the correct area for other question types, the model, in the "relation" map indicating the correct question type, focused on the obscure person in the background rather than the visible girl in the image, yet still produced the correct answer. In 21 of the questions, the model focused on the relevant parts of the image but gave the wrong answer. In 11 of the questions, the model focused on irrelevant parts and predicted the wrong answer, *e.g.*, Question 2 in Figure 3. The model's tendency to focus on irrelevant areas in a substantial portion of correct answers suggests that instead of true reasoning, the model might be relying on learned correlations or biases within the dataset to make informed guesses.

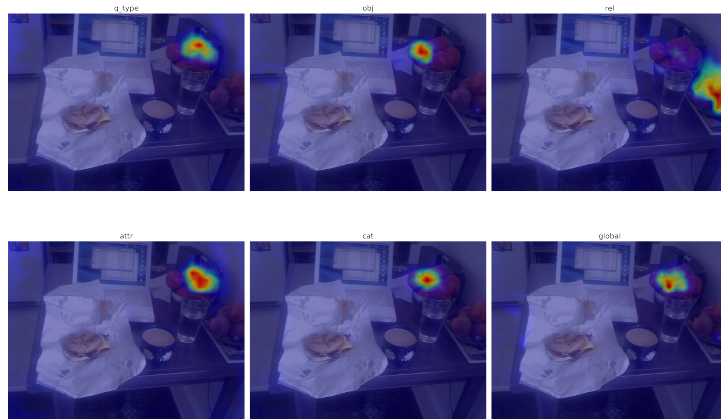
While attention visualization offers insights into the model's reasoning mechanism through local explanations, it may be inadequate for fully understanding the reasoning processes of MLLMs, particularly those with billions of parameters. In recent years, approaches like Chain of Thought (CoT) [18, 19] have been developed to both trigger reasoning mechanisms and make reasoning more transparent in LLMs/MLLMs. In this thesis, the goal is not only to enhance the VQA performance but also to increase the transparency of reasoning by converting images into context-aware textual descriptions and producing rationales for question-answering.

### 3.3 Methodology

#### 3.3.1 Image Captioning-Integrated VQA Pipeline

This study explores the effectiveness of replacing images with descriptive captions in the VQA pipeline and utilizing LLMs as question-answering models in zero-shot settings. The main hypothesis of this study is that context-aware image captioning could act as a better intermediary between images and LLMs.

Several studies [83, 84, 85] explore integrating image captions into the VQA pipeline using LLMs. Our study diverges from these studies in several aspects. Firstly, this study proposes a new approach to leverage image captions in the VQA pipeline using the context of the question, called question-driven image captioning. Also, the study conducts a thorough evaluation of various image captioning and question-answering models in the VQA pipeline across various questions in terms of semantics and structure.



**Question 1:** Are the tomatoes round and red? **Type:** attr **Answer:** yes **Prediction:** yes



**Question 2:** What vehicle is not red? **Type:** cat **Answer:** ambulance **Prediction:** truck



**Question 3:** Who is wearing the shirt? **Type:** rel **Answer:** girl **Prediction:** girl

Figure 3: Visualization of encoder-decoder multi-head attention weights



Therefore, a traditional VQA pipeline has been rethought as a combination of image captioning (IC) and question-answering (QA) phases. For the remainder of the study, this pipeline is defined as the IC-integrated VQA pipeline. In the IC-integrated VQA pipeline, an image captioning model first transforms images into informative captions. These captions are then processed by an LLM to generate answers. To improve this pipeline, a new, straightforward, context-aware approach called question-driven image captioning is introduced in this study. Throughout this thesis, the proposed image captioning approach for VQA has been referred to as question-driven image captioning (QD-IC) or context-aware image captioning.

The proposed approach incorporates contextual information from the question into the IC model through the extracted keywords. This method enables the IC model to generate captions that focus solely on the relevant parts of the image, customizing the captions according to the specific question. The proposed VQA pipeline is given in Figure 4.

The performance of the IC-integrated VQA pipeline was evaluated using various IC models and compared across different semantic and structural question categories comprising the GQA dataset.

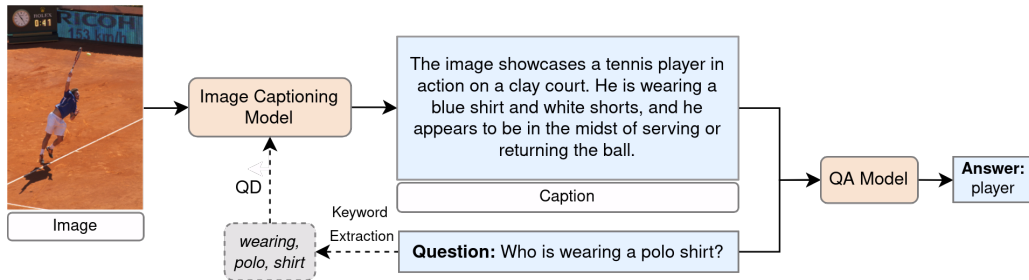


Figure 4: VQA pipeline exploiting general and the proposed question-driven image captioning (QD-IC) as an intermediate step.

### 3.3.1.1 Vision-Language Models for Caption Generation

The primary stage of the proposed pipeline is image captioning (IC). The IC stage is significant for performing the question-answering (QA) phase with relevant and informative content. Image captions serve as condensed summaries of the visual content, yet there is a risk of losing specific visual details that could affect the QA performance. Hence, a review of IC models was conducted, focusing on the models providing detailed captions while considering our computational constraints. Vision-language models indicating high performance in IC benchmarks, including CogVLM [12], FuseCap [86], and BLIP-2 OPT<sub>2.7b</sub> [13], and recent MLLMs GPT-4 [8] and GPT-4o [74] were selected, and incorporated as caption generators into the VQA pipeline. The IC models were provided from HuggingFace and open-access GitHub repositories. Only pre-trained weights provided were utilized, and no additional training was performed. An overview of the selected models is given in the following paragraphs.

**CogVLM** [12] is an open-source, pre-trained vision-language model featuring 10B visual and 7B language parameters. CogVLM integrates a trainable visual expert into the language model, enabling each layer to leverage a new Query-Key-Value matrix and Multilayer Perceptron (MLP) layer to combine image and text features. CogVLM-17B provides different downstream models among which chat

and visual grounding variants were utilized in this study. The performance of the CogVLM-chat<sup>3</sup> and CogVLM-visual grounding<sup>4</sup> models was assessed according to various categories of questions. Additionally, 4-bit quantization was applied to the CogVLM models to align with our computational constraints.

**BLIP-2 OPT<sub>2.7b</sub>** [13] contains a pre-trained Querying Transformer (Q-Former) designed to leverage a set of learnable query vectors for extracting visual features from a frozen image encoder. The language model in this variant of BLIP-2 is OPT [87], a decoder-based LLM with 2.7B parameters. Specifically, pre-trained BLIP-2 OPT<sub>2.7b</sub><sup>5</sup> was exploited. BLIP-2 was executed in F16 precision to fit within our computational resources.

**FuseCap** [86] uses a set of visual experts to augment existing image captions with more visual details. For this purpose, it uses a BLIP-based vision-language model to generate base captions. Then, the original caption and the outputs of several visual experts for object detection, attribute recognition, and Optical Character Recognition (OCR) are fused by an LLM for a more descriptive caption generation. Specifically, the open-access FuseCap<sup>6</sup> model was utilized.

**GPT-4** [8] and **GPT-4o** [74] are the recent MLLMs released by OpenAI in 2023 and 2024, respectively. GPT-4 supports vision and language modalities, while GPT-4o also supports audio inputs. Technical details of the models are not published. They are accessible only through API calls.

### 3.3.1.2 Caption Generation Process

Image captions were generated in two different ways. In the first approach, referred to as "general-purpose image captioning (GP-IC)", the model takes a general prompt and the image as input to describe the scene. Although a general-purpose image caption provides a broad understanding of the visual content, it may not contain the exact information needed to address the key aspects of the associated question. To address this limitation, this study proposes a new approach called "question-driven image captioning (QD-IC)" to provide contextually relevant captions for image-question pairs. In this approach, the image captioning model utilizes both the image and context keywords embedded in the prompt as inputs.

KeyBERT [88] was utilized to extract contextual keywords from the questions. KeyBERT first employs BERT to extract document embeddings for a document-level representation. Then, word embeddings are generated for N-gram words/phrases. Using cosine similarity, the words/phrases most similar to the document are identified as the best descriptors of the document. KeyBERT was provided with each question in the test set and the most representative keywords (unigrams) were extracted. The extracted keywords were then embedded into the prompt and fed into the IC model along with the related image. Table 3 provides examples of questions, with keywords underlined, alongside the corresponding image captions.

During the experiments, initially, the BLIP-2, FuseCap, CogVLM-chat (CogVLM-C), CogVLM-visual grounding (CogVLM-V), GPT-4, and GPT-4o models were employed for general-purpose im-

---

<sup>3</sup> <https://huggingface.co/THUDM/cogvlm-chat-hf>

<sup>4</sup> <https://huggingface.co/THUDM/cogvlm-grounding-generalist-hf>

<sup>5</sup> <https://huggingface.co/Salesforce/blip2-opt-2.7b>

<sup>6</sup> <https://github.com/RotsteinNoam/FuseCap>

Table 3: Comparison of IC-integrated and VQA pipelines evaluated using GPT-4o.

Question-Keywords	Image Caption
What is the <u>man</u> to the <u>left</u> of the <u>glasses</u> <u>doing</u> ?	In the image, there are three people sitting on a bench facing the beach and the ocean. The person on the left, who appears to be a man, is wearing glasses and a cap. The people are shaded by two umbrellas, one pink and one white with a floral pattern. The man on the left is sitting and looking towards the ocean, seemingly enjoying the view. The scene is calm and serene, with the beach and ocean in the background under a clear blue sky.
The <u>parked vehicles</u> are <u>waiting</u> for what?	The image shows a traffic light at an intersection, displaying a red light. Several vehicles are waiting at the intersection, indicating they are stopped due to the red light. In the background, there are residential buildings with multiple floors. The road appears to be relatively empty, with only a few vehicles visible. The scene is set in an urban area with a mix of residential and road infrastructure.
Are the <u>shorts large</u> and <u>blue</u> ?	The image shows a person standing on a sandy beach near the ocean. They are holding a large surfboard with a pink design featuring flowers. The person is wearing a red shirt and blue shorts. The beach is surrounded by hills and the sky appears to be overcast. There is also a large white banner with blue text partially visible in the foreground.

age captioning. Subsequently, the highest-performing models, CogVLM-C, GPT-4, and GPT-4o, were utilized for question-driven image captioning. These experiments investigate the differences between general-purpose image captioning and question-driven image captioning, as well as evaluate the effectiveness of captions generated by various vision-language models in the question-answering phase.

Furthermore, the study examines whether irrelevant portions of an image caption could cause confusion or result in inaccurate answers from the QA model. Therefore, an experiment was conducted to determine the impact of providing only the most relevant sentence of the image caption to the LLM during the QA phase. For this purpose, Sentence-BERT [89], specifically the MiniLM-L6 model<sup>7</sup>, was used to extract the most pertinent sentence from the image caption based on the given question. The extracted sentence, along with the question, was then provided to the LLM in the QA phase.

### 3.3.1.3 Question Answering

In the proposed pipeline, the QA model is an LLM taking context-aware image captions and questions as input and generating answers accordingly. Recently, LLMs have become state-of-the-art in various language tasks, including text generation, translation, and question-answering. Their performance and generalization capabilities make them effective for language processing tasks. Therefore, this study employs LLMs during the QA phase due to their superior performance in QA benchmarks [90].

During the QA phase, OpenAI’s GPT-3.5, GPT-4, and GPT-4o models were utilized. The majority of experiments were conducted using GPT-3.5 and GPT-4o due to their cost-effectiveness. Answers were generated through OpenAI’s GPT APIs, with optimal performance achieved using the following configuration. The *temperature* parameter was set to 0.1 to minimize creativity and ensure more consistent results. Considering that the answer might be within the image caption, the *presence penalty* was set

<sup>7</sup> <https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1>

to 0 to avoid penalizing repeated words in the prompt (image captions and questions). To minimize the repetition of a word in the answer, the *frequency penalty* was also set to 0. The final configuration of these parameters was determined through manual observations.

### 3.3.1.4 Baseline VQA Models

To compare the performance differences between a standard VQA approach with the IC-integrated VQA approach, several vision-language models (or MLLMs) were exploited as baseline VQA models. The chat variation of CogVLM<sup>8</sup> and BLIP-2 FlanT5<sub>XL</sub><sup>9</sup> were employed to perform VQA. CogVLM surpasses many other vision-language models, including InstructBLIP [91] and LLaVA-1.5 [92], in VQA benchmarks. Similarly, BLIP-2 FlanT5<sub>XL</sub>, with 4.1B parameters, demonstrates strong performance, exceeding BLIP-2 OPT<sub>2.7b</sub> and Flamingo [11] in VQA benchmarks. Due to resource constraints with 16 GB VRAM, CogVLM and BLIP-2 FlanT5<sub>XL</sub> were utilized with 4-bit quantization and F16 precision, respectively. Additionally, APIs for the latest MLLMs, GPT-4 and GPT-4o by OpenAI, were leveraged to perform VQA.

### 3.3.1.5 Prompt Details

To generate general-purpose image captions, the prompt "*Describe the scene in this image*" was given to the CogVLM. To create question-driven image captions, the prompt was adapted to include "*Consider the keywords: [keywords]*".

QD-IC was also conducted using GPT-4o. A similar prompt was fed into GPT-4o, incorporating the image, context keywords, and instruction, as shown in Figure 5.

Prompt for Image Captioning with GPT-4o
<p><b>System:</b> You are a helpful assistant describing the image focusing on the given context keywords.</p> <p><b>User:</b> [Keywords, Image]</p>

Figure 5: Prompt for QD-IC with GPT-4

During the open-ended answer generation in QA, the number of words produced was limited to 2 to align with the length of the ground-truth answers in the GQA dataset. The prompt fed into GPT-3.5 for QA is shown in Figure 6. It was observed that including an instruction to consider the question type in the prompt helps obtain the expected output, *e.g.*, "yes/no" for existence questions.

During the prompt engineering process for the QA step with GPT-4o, observations revealed that GPT-4o often provides answers with specific details. To address this, an instruction was added to the prompt to encourage broader responses rather than specific ones, as shown in Figure 7. Another issue encountered was that GPT-4o often provided responses other than "yes" or "no" to binary questions. Therefore, an instruction was added to the prompt to emphasize the need for binary responses.

<sup>8</sup> <https://huggingface.co/THUDM/cogvlm-chat-hf>

<sup>9</sup> <https://huggingface.co/Salesforce/blip2-flan-t5-xl>

<b>Prompt for Question Answering (GPT-3.5)</b>
<p><b>System:</b> You are a helpful assistant answering the question in only one or two words based on the 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.</p> <p><b>User:</b> [Caption, Question]</p>

Figure 6: Prompt for QA with GPT-3.5

<b>Prompt for Question Answering (GPT-4o)</b>
<p><b>System:</b> You are a helpful assistant answering the question in only one or two words based on the text. A binary question should only be answered with 'yes' or 'no'. Provide answers that are broad and general rather than specific.</p> <p><b>User:</b> [Caption, Question]</p>

Figure 7: Prompt for QA with GPT-4o

To evaluate the performance of the IC-integrated VQA pipeline in comparison to MLLMs such as CogVLM, BLIP-2, GPT-4, and GPT-4o, standard VQA was performed utilizing the prompt depicted in Figure 8. Experiments with different prompt strategies for VQA are provided in Section 3.6.

<b>Standard Prompt for VQA</b>
<p><b>System:</b> You are a helpful assistant answering the questions in one or two words based on the image content.</p> <p><b>User:</b> [Question, Image]</p>

Figure 8: Prompt for VQA with MLLMs

### 3.4 Results

Table 4 presents the evaluation results for the IC-integrated VQA pipeline containing various vision-language models for image captioning and LLMs for question-answering. In the evaluation, as detailed in Section 3.1.3, a model’s prediction was considered correct if the cosine similarity ratio between the label and the prediction exceeded the specified threshold, and accuracy measurement was conducted accordingly. Table 4 compares the results at cosine similarity threshold 0.7. Results indicate that question-driven image captioning (QD-IC) improves the accuracy in most question categories compared to general-purpose image captioning (GP-IC). Results at threshold 0.8 and exact matching are given in Appendix B.

Table 4: Comparison of the accuracy values of various IC-integrated VQA pipelines at cosine similarity 0.7, containing the proposed question-driven image captioning (QD-IC) and general-purpose image captioning (GP-IC) on GQA *test-dev-balanced*.

Question type	CogVLM-C GP-IC + GPT-3.5 QA	CogVLM-V GP-IC + GPT-3.5 QA	CogVLM-C QD-IC + GPT-3.5 QA	CogVLM-C QD-IC + GPT-4o QA	FuseCap GP-IC + GPT-3.5 QA	BLIP-2 GP-IC + GPT-3.5 QA	GPT-4 QD-IC + GPT-4o QA	GPT-4o QD-IC + GPT-4o QA	GPT-4o GP-IC + GPT-4o QA
verify	63.01	58.53	66.83	64.87	53.60	55.82	69.58	<b>74.33</b>	64.87
query	36.91	31.08	38.34	35.53	29.61	31.87	36.84	<b>40.08</b>	35.59
choose	65.25	60.90	65.51	59.30	58.07	60.82	61.17	<b>72.42</b>	54.43
logical	59.51	60.29	59.07	63.72	57.07	56.07	68.16	<b>71.32</b>	64.78
compare	51.78	51.95	51.95	52.63	54.50	48.22	53.98	<b>55.68</b>	48.55
object	61.95	63.24	59.13	68.63	59.38	58.35	71.20	<b>75.44</b>	72.23
attribute	51.75	46.42	54.62	47.98	45.11	46.63	54.25	<b>58.10</b>	45.49
category	47.35	44.21	50.39	51.08	43.52	42.47	49.08	<b>53.87</b>	52.30
relation	42.56	38.32	42.97	43.89	34.98	37.23	43.65	<b>47.96</b>	44.32
global	<b>49.04</b>	45.86	45.86	43.31	43.95	45.22	40.12	43.31	48.40
total	48.06	43.83	49.50	47.75	41.58	42.99	50.23	<b>54.35</b>	47.33

### 3.4.1 Usefulness of Question-Driven Image Captions

Applying the question-driven image captioning (QD-IC) approach to the models CogVLM-C and GPT-4o, which demonstrated the best VQA performance with general-purpose image captioning (GP-IC), resulted in an overall improvement in VQA accuracy.

Using CogVLM for image captioning, the QD-IC approach led to improvements in all structural question categories compared to the GP-IC approach, except for the *logical* category. These improvements are particularly pronounced in the *verify* category which consists of "yes/no" questions, with a 3.8% increase. Accuracy improvements among semantic classes are especially notable in *category* and *attribute* types, with a 3% increase. The questions in these categories are mostly related to the identification or location of a specific object or its attributes in the image. Overall, the greatest increase in correct answers was observed in questions belonging to the *verify-attribute* structural-semantic category. These "yes/no" questions typically ask whether an object in the scene possesses a certain attribute, such as "*Does the clay flower pot look blue?*". Examples of correct predictions from the mentioned question categories are provided in Figure 9.

The highest performance among IC-integrated VQA pipelines was achieved with the *GPT-4o QD-IC + GPT-4o QA* configuration, where GPT-4o performed question-answering on QD image captions generated by GPT-4o. The question-driven image captioning applied with GPT-4o increased overall accuracy by 7.02% compared to general-purpose image captioning. The most significant performance improvement was observed in the *choose* category, with a 17.99% increase. The *choose* category pertains to questions asking which of the alternatives is valid for an object in the image, such as "*Does the shirt seem to be sleeveless or long-sleeved?*". In this category, context words extracted from the question include the main subject of the sentence and the alternatives being asked for chosen. Therefore, QD image captions are usually oriented on the correct visual details and help ease the question-answering process.

Other notable accuracy improvements with GPT-4o were observed in the *verify* category with a 9.46% increase, the *attribute* category with a 12.61% increase, and the *compare* category with a 7.13% increase. Analyses suggest that the general reason for the increase in most question categories is that, in cases where the visual details sought in the question are not the main subject of the image, the question-driven image captioning approach draws attention to the keywords, thereby generating a caption that includes the asked visual details. Corresponding examples can be seen in Figure 9.

A performance drop with CogVLM in the QD-IC integrated pipeline has been observed in the *object* category. However, a closer look shows that inaccuracies arise due to the behavior of GPT-3.5 during the QA, phase despite the QD image captions containing relevant information. This issue will be discussed in the Section 3.4.4.

It was also observed that the focus of QD-IC was associated with a performance decline with both CogVLM-C and GPT-4o in the *global* category, which typically involves questions about the overall content of an image. This decline is primarily due to the limited availability of keywords that can provide adequate context for *global* questions. The limitation of context keywords prevents the context from being effectively conveyed to the image captioning model, often leading to captions that are not sufficiently informative for questions about the entire scene. Consequently, the QD-IC approach has had a negative impact on this type of question. Moreover, a significant portion of the predictions, though not perfectly aligned with the labels, have still been considered acceptable. This issue has been particularly prevalent in the *global* category, where questions pertain to the entire scene.

### 3.4.2 Effect of Image Captioning Model

By utilizing the question-driven image captioning approach to generate more descriptive captions, GPT-4o outperformed all other IC models, including CogVLM, GPT-4, FuseCap, and BLIP-2, significantly enhancing answer accuracy across most question categories. With GPT-4o fixed as the QA model, performance was evaluated using QD-IC image captions generated by CogVLM-C, GPT-4, and GPT-4o. The results showed that GPT-4o’s image captions led to an overall accuracy improvement of 6.6% over those from CogVLM and 4.12% over those from GPT-4. It was observed that GPT-4o, particularly when compared to the closely performing GPT-4, generated more informative captions in specific question categories. Notably, there was an 11.25% increase in the *choose* category, a 4.79% increase in the *category* category, and a 4.75% increase in the *verify* category. Correct predictions made by the *GPT-4o QD-IC + GPT-4o QA* configuration are illustrated in Figure 11.

Another performance difference was observed between the CogVLM-chat (CogVLM-C) and visual grounding (CogVLM-V) variants in general-purpose image captioning. Overall, CogVLM-C demonstrated better performance than CogVLM-V. However, CogVLM-V led to a notable performance improvement in the *object* semantic question category. This suggests that the *object* category involves questions that are closely related to object detection and visual grounding tasks, e.g., "*Is there any bag that is black?*", using a visual grounding model as the base improves the performance in this specific question category.

Question: What kind of furniture is wooden?  
 Label: table  
 Prediction: table  
 Semantic\_type: cat  
 Structural\_type: query  
 Text: The image showcases a breakfast setting on a wooden table. The main dish is a large pancake topped with a dollop of whipped cream. To the side, there's a serving of scrambled eggs and a strip of bacon. The overall ambiance seems cozy and inviting, with a touch of elegance brought by the whipped cream on the pancake.



Question: Is the mug in front of the cup green and small?  
 Label: no  
 Prediction: no  
 Semantic\_type: attr  
 Structural\_type: logical  
 Text: The scene in the image showcases a table set for a formal event. There are white cups and mugs, a vibrant green napkin, and a centerpiece with flowers. The ambiance suggests a sophisticated gathering, possibly a dinner or a gala.



Question: Is there a blender to the right of the yellow drink?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The scene in the image showcases two individuals outdoors, with one person pouring a liquid from a bottle into a blender placed on top of yellow crates. The setting appears to be a casual gathering or event, with various items scattered around, including bicycles and a warning sign.



Figure 9: Examples of accurate predictions by *CogVLM QD-IC + GPT-3.5 QA*

### 3.4.3 Effect of Filtering Image Captions

During the experiments, the effect of filtering image captions by extracting only the most relevant sentence, referred to as sentence-based image captioning (SB-IC), was also evaluated. The results indicate that SB-IC, with an accuracy of 44.12%, underperformed in all question categories compared to methods that did not narrow down the captions. This suggests that limiting the information in the captions and providing shorter inputs to the QA model is ineffective, leading to a loss of crucial information and a consequent decline in performance.

### 3.4.4 Effect of the Question-Answering Model

When using identical image captions (CogVLM-C QD-IC) within the pipeline, notable differences in QA performance between GPT-3.5 and GPT-4o were observed. Despite GPT-4o being a significantly larger model and outperforming GPT-3.5 across numerous benchmarks, it exhibited weaker performance on CogVLM-C image captions during the QA phase.

Throughout the experiments, it was observed that GPT-4o tended to provide more specific answers to the questions and frequently responded with "unknown" when it could not derive a specific answer from the image caption. In contrast, GPT-3.5 provided broader answers. For example, GPT-3.5's answer to the question "How tall is the chair at the bottom of the photo?" was *standard height*, whereas GPT-4o's response was *unknown*.

Additionally, GPT-3.5 demonstrated a tendency to utilize external knowledge beyond the provided text to answer questions and make informed guesses. For instance, while GPT-4o answered the question "What are the glasses made of?" with *unknown*, GPT-3.5 responded with *glass*, even though the image caption does not include this information. Similarly, for the question "What is the phone made of?",



GPT-3.5’s answer was *plastic*, which is the correct answer, while GPT-4o’s response was *unknown*. This performance disparity between the two models is thought to stem from this behavioral difference. This issue will be further explored in the error analysis in Section 3.5.

Utilizing GPT-4o as the QA model significantly boosted performance in the *object* category, with a 9.5% increase in accuracy. Examples of incorrect predictions of GPT-3.5 on CogVLM-C QD image captions in the *object* category are shown in Figure 10. As shown in the figure, a prominent error in the *object* category is responding to a "yes/no" question asking whether one of two objects is present in the scene with the object’s name. The correct response should simply be *yes* if one of the objects is indeed present. For example, in Figure 10a, GPT-3.5 incorrectly answered *women* to the question "Are there glasses or women?" instead of the correct response, *yes*.

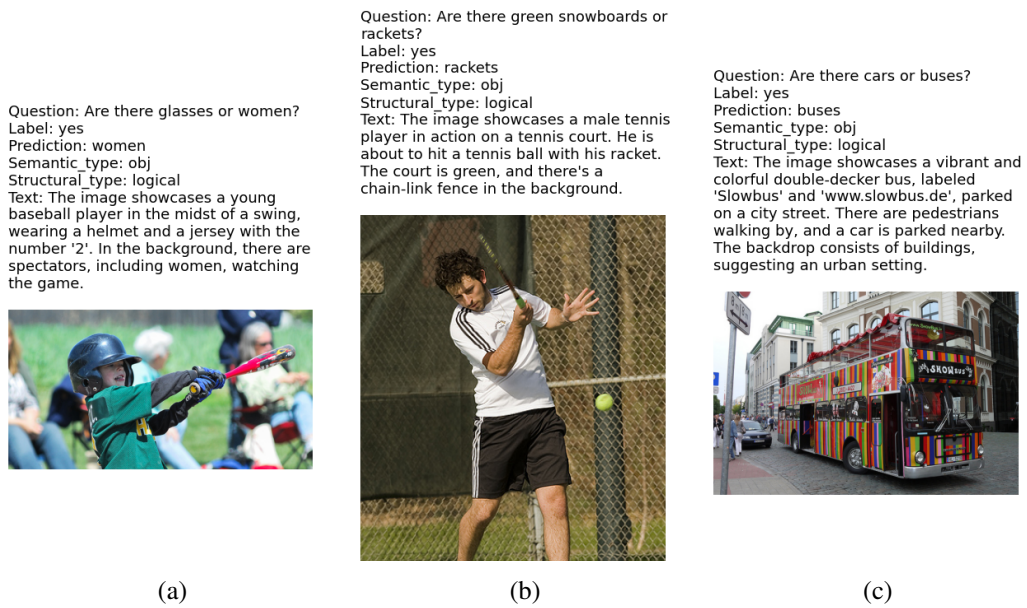


Figure 10: Examples of incorrect predictions of *GPT-3.5* in the *object* question category during the QA on CogVLM-C QD image captions. It is observed that GPT-4o correctly answers the same questions.

### 3.4.5 Predictability of Answers in the Test Set

To assess the predictability of answers in the GQA test set without visual input, the GPT-3.5 was employed to predict answers based solely on the provided questions. The prompt provided was: "You are an assistant, predicting the most probable answer for a given question, and the answer must be one word only.". It was observed that 3,745 of the answers predicted by GPT-3.5 matched the ground truth answers, resulting in an accuracy of 29.77%. Notably, 2,099 of these correct answers, accounting for 56%, were in the form of "yes/no" responses. These results indicate that GPT-3.5 can predict a significant portion of the questions even without visual input.

Question: What animal is the couch behind of?  
 Label: dog  
 Prediction: dog  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image depicts a festive indoor scene with a decorated Christmas tree adorned with gold and blue ornaments and tinsel. To the left of the tree, there is a snowman decoration. In front of the tree, there are two small reindeer figurines. To the right, a fluffy white dog is wearing a red Santa hat and is sitting on a couch. The background includes a framed picture on the wall and some plush toys on the couch. The overall atmosphere is cheerful and festive, celebrating the holiday season.



Question: Is the field soft and snowy?  
 Label: no  
 Prediction: no  
 Semantic\_type: attr  
 Structural\_type: logical  
 Text: The image shows a child playing baseball on a grassy field. The child is wearing a helmet and a red hoodie and is swinging a bat at a baseball on a tee. There are several baseballs scattered on the grass, and a bag with a baseball glove is lying on the ground nearby. The field is green and does not appear to be snowy. The scene does not convey a snowy or soft field.



Question: Is that shoe behind a dog?  
 Label: no  
 Prediction: no  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image shows a Siamese cat next to a pair of black and white sneakers. There is no dog visible in the image. The cat is positioned close to the shoes, and a person is partially visible in the background. The scene appears to be outdoors, with some greenery, including a fern, in the background.



Question: What is waiting for the traffic light?  
 Label: cars  
 Prediction: vehicles  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a traffic light at an intersection. The traffic light is red, indicating that vehicles must stop and wait. In the background, there are several apartment buildings and a few cars are visible, some of which are waiting at the intersection. The road is relatively empty, and the area appears to be a residential neighborhood. The sky is overcast, suggesting it might be a cloudy day.



Question: How is the clothing item that is pink called?  
 Label: tank top  
 Prediction: tank top  
 Semantic\_type: cat  
 Structural\_type: query  
 Text: The image shows a person playing tennis. They are wearing a pink tank top, which is the clothing item referred to.



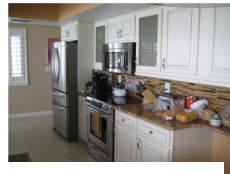
Question: How wide is the parking lot made of cement?  
 Label: wide  
 Prediction: wide  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image depicts a wide cement parking lot with a person performing a skateboard trick. The parking lot is empty, with clearly marked parking spaces. The person is in mid-air, seemingly executing a jump or flip with the skateboard. The background features a line of trees and bushes, adding a natural element to the scene.



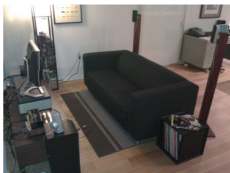
Question: Is there an elephant near the person that is wearing a coat?  
 Label: no  
 Prediction: no  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image shows two people dressed in winter clothing, including coats, in a snowy environment. One person is wearing a pink and purple coat and is holding ski poles, while the other person is wearing a white and gray coat and is standing nearby. There is a sign in the background with a silhouette of a person skiing. There is no elephant in the image.



Question: Are there refrigerators to the left of the stove?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image shows a modern kitchen with a focus on the appliances and cabinetry. On the left side of the image, there is a stainless steel refrigerator. Next to the refrigerator, there is a stainless steel stove with an oven, and above the stove, there is a microwave. The kitchen cabinets are white with a classic design, and the countertop is a brownish granite. The backsplash is a mosaic of brown, beige, and white tiles. Various kitchen items, such as a toaster, a coffee maker, and some food items, are placed on the countertop.



Question: What is the picture frame hanging from?  
 Label: wall  
 Prediction: wall  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image depicts a living room or a media room setup. There is a black couch positioned on a rug in the center of the room. In front of the couch, there is a TV on a stand with various electronic devices and cables around it. To the left of the couch, there is a tall lamp and a framed picture hanging on the wall. The picture frame is positioned above a shelf or a stand with some items on it. In the background, there is a desk with a chair and some office equipment. The room has a wooden floor and appears to be well-organized.



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 shows a green bag lying on a bed with various items spread out around it. The items include notebooks, pens, a hairbrush, a phone, a wallet, a bottle of lotion, a tube of lip balm, a pair of sunglasses, a small bottle of hand sanitizer, a compact mirror, a set of keys, and other personal items. The bed has a white bedspread with a subtle floral pattern. There is no leather sofa visible in the image.



Question: What device is above the keyboard?  
 Label: monitor  
 Prediction: monitor  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a desk setup featuring a keyboard and a computer monitor. The keyboard is a white mechanical keyboard with a standard layout. The monitor is a CRT (cathode-ray tube) display. Next to the keyboard, there is a computer mouse and a mouse pad. The desk also has a few other items, including a water bottle, a black mug, and some papers. The setup appears to be in an office or a home office environment.



Question: What is the color of the glove?  
 Label: black  
 Prediction: black  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image shows a baseball game in progress. The batter, wearing a navy blue jersey with the number 13, is standing in the batter's box. The batter's gloves are black. The catcher, positioned behind the batter, is wearing a dark green jersey. The scene is viewed through a chain-link fence.



Figure 11: Examples of correct predictions by GPT-4o QD-IC + GPT-4o QA

### 3.4.6 Comparison with the Baseline VQA Models

Table 5 compares the best-performing IC-integrated VQA pipelines with the standard VQA performance of various MLLMs under zero-shot settings. The results show that the *GPT-4o QD-IC + GPT-4o QA* pipeline, which is the best among the IC-integrated VQA approaches, does not match the standard VQA performance of CogVLM-C. However, generating image captions with GPT-4o and then using them for QA with GPT-4o, rather than directly providing images and questions, led to improvements in some question categories, albeit at the cost of overall performance.

The most notable improvement was observed in the *object* category, with a 9.24% increase in accuracy. The *logical* and *compare* categories also showed accuracy gains of 2.42% and 2.88%, respectively. In contrast, the largest declines in accuracy were noted in the *choose* and *attribute* categories. A bar chart comparing the correct predictions across different question categories is provided in Appendix B.2. Overall, IC-integrated pipelines with CogVLM-C and GPT-4o outperform BLIP-2 FlanT5<sub>XL</sub> in zero-shot settings across most question categories. More examples of correct and incorrect predictions of evaluated models are given in Appendix A.

Table 6 presents a comparison of the overall performance between IC-integrated and standard VQA models across different cosine similarity thresholds, Exact Match (EM) rates and the metric Union-AAS [78]. The best performance among IC-integrated approaches for all the metrics was achieved with *GPT-4o QD-IC + GPT-4o QA*. Although this performance surpasses the VQA performance of BLIP-2 and GPT-4, it remains below the overall VQA performance of GPT-4o and CogVLM. Union-AAS metric was measured on the alternative answer set (AAS) [78] providing acceptable answers for each ground truth answer in the *GQA test-dev-balanced*. Measurements conducted using this metric were found to be consistent with those obtained from other metrics.

Overall, standard VQA with MLLMs surpasses the best-performing IC-integrated VQA pipeline. These findings highlight the limitations of current image captioning models in fully capturing and representing the various critical aspects of a scene, even when provided with context-aware keywords. However, the IC-integrated VQA approach has shown potential in improving performance on questions involving object and attribute identification, especially when visual details pertain to objects that are not the main focus of the image. Future research could enhance context keywords by making them more informative and by utilizing bigrams and trigrams instead of unigrams.

### 3.4.7 VQA Evaluation using LLMs

The primary issue with the evaluation is that LLMs generate open-ended responses, which do not always align with the ground truth. To account for synonyms and very similar words, such as *sofa* and *couch*, *woman* and *lady*, or *dark* and *black*, previous evaluations used an accuracy metric incorporating word similarity scores. However, this method still fails to capture other valid answers that deviate from the exact ground truth. For instance, responses like *man on skateboard* or *man* to the question "*The woman to the right of the camera is watching who?*", associated with the ground-truth *skateboarder*, were not accepted correct in the prior evaluation.

To address this issue, as an alternative evaluation approach, predictions were assessed automatically using a powerful LLM, GPT-4o. In this approach, GPT-4o was provided with the question, the ground truth answer, and model predictions, and tasked with scoring the predictions on a scale of 0-2 con-

Table 5: Comparison of the performances of VQA and QD-IC + QA pipelines

Question type	CogVLM-C VQA	BLIP-2 VQA	GPT-4 VQA	GPT-4o VQA	GPT-4 QD-IC + GPT-4o QA	GPT-4o QD-IC + GPT-4o QA	CogVLM-C QD-IC + GPT-4o QA
verify	<b>83.04</b>	56.48	71.14	78.91	69.58	74.33	64.87
query	<b>54.11</b>	41.31	46.19	50.99	36.84	40.08	35.53
choose	<b>87.32</b>	56.91	74.82	83.51	61.17	72.42	59.30
logical	<b>77.54</b>	54.24	45.65	68.89	68.16	71.32	63.72
compare	<b>62.65</b>	46.52	55.86	52.80	53.98	55.68	52.63
object	<b>84.45</b>	57.07	46.27	66.20	71.20	75.44	68.63
attribute	<b>70.45</b>	49.33	57.78	69.16	54.25	58.10	47.98
category	<b>63.19</b>	53.35	53.35	59.62	49.08	53.87	51.08
relation	<b>59.91</b>	43.31	51.15	54.41	43.65	47.96	43.89
global	<b>56.05</b>	40.13	36.31	43.31	40.12	43.31	43.32
total	<b>66.02</b>	47.52	53.61	61.58	50.23	54.35	47.75

Table 6: Comparison of overall accuracy in IC-integrated VQA pipelines and conventional VQA pipelines for exact matching (EM) and different cosine similarity thresholds.

Models	EM	sim=0.9	sim=0.8	sim=0.7	Union-AAS
CogVLM-C GP-IC + GPT-3.5 QA	36.77	38.21	43.01	48.06	-
CogVLM-V GP-IC + GPT-3.5 QA	36.21	37.51	41.21	43.83	-
CogVLM-C SB-IC + GPT-3.5 QA	34.14	35.06	39.41	-	-
CogVLM-C QD-IC + GPT-3.5 QA	37.64	39.24	44.48	49.50	39.86
CogVLM-C QD-IC + GPT-4o QA	42.10	42.90	45.22	47.75	45.16
FuseCap GP-IC + GPT-3.5 QA	33.17	34.18	37.64	41.58	34.80
BLIP-2 GP-IC + GPT-3.5 QA	34.77	35.53	39.11	42.99	36.47
GPT-4 QD-IC + GPT-4 QA	35.86	39.32	43.75	-	38.14
GPT-4 QD-IC + GPT-4o QA	43.63	44.71	47.40	50.23	46.51
GPT-4o QD-IC + GPT-4o QA	<b>47.90</b>	<b>49.14</b>	<b>51.67</b>	<b>54.35</b>	<b>51.08</b>
CogVLM VQA	<b>58.43</b>	<b>59.23</b>	<b>62.79</b>	<b>66.02</b>	<b>62.71</b>
GPT-4o VQA	54.03	55.32	58.82	61.58	57.30
BLIP-2 VQA	37.82	38.57	42.33	47.52	40.17
GPT-4 VQA	42.02	45.26	49.55	53.61	44.97

sidering the question and the ground truth. A score of 2 indicates correct or acceptable answers, 1 indicates ambiguous answers, and 0 indicates incorrect or irrelevant answers. The prompt in Figure 12 was designed with the inspiration of a recent study [93]. This method of VQA evaluation can be further enhanced by incorporating images as inputs in addition to predictions and ground truth answers to better determine the plausibility of the answers.

The results of the LLM evaluation are summarized in Table 7. These findings align with existing evaluations, which indicate CogVLM as the highest-performing VQA model. The LLM evaluation appeared effective in capturing close-meaning words and phrases. For instance, *sitting* is considered an acceptable answer to the question "What is the man to the left of the glasses doing?", which has the ground-truth answer *resting*. Ambiguous results typically arise when the ground truth and prediction

Prompt for VQA Evaluation using GPT-4o
<p><b>System:</b> You are a helpful assistant evaluating the predicted answers of the models to a question about an image. Your task is to rate the accuracy of each predicted answer on a scale of 0-2.</p> <p>Use the following scoring:</p> <ul style="list-style-type: none"> <li>- 0 indicates incorrect or irrelevant answers.</li> <li>- 1 indicates ambiguous answers.</li> <li>- 2 indicates correct or acceptable answers.</li> </ul> <p>VERY IMPORTANT:</p> <ul style="list-style-type: none"> <li>- If the answer matches the ground-truth, rate it as 2.</li> <li>- If the answer is a synonym or plural/singular form of the ground-truth, rate it as 2.</li> <li>- For binary questions, only 'yes' or 'no' are plausible answers. Any other response should be rated as 0.</li> </ul> <p><b>User:</b> [Question, Model Predictions, Ground-truth]</p>

Figure 12: Prompt for VQA Evaluation using LLMs

do not exactly match or are interchangeable in the context of the question, but share overlapping aspects in meaning. For example, the prediction *beach* for the question "What is the airplane flying above?" with the ground-truth answer *ocean* was classified as ambiguous by the LLM. Manual observations suggest that LLM evaluation may improve the judgment of predictions in open-ended generation tasks.

Table 7: Comparison of IC-integrated and VQA pipelines evaluated using GPT-4o.

Models	Correct (2)	Ambiguous (1)	Incorrect (0)
CogVLM-C QD-IC + GPT-3.5 QA	48.38	9.89	41.74
CogVLM-C QD-IC + GPT-4o QA	47.84	9.16	43.00
FuseCap GP-IC + GPT-3.5 QA	37.95	6.45	55.60
GPT-4 QD-IC + GPT-4 QA	42.93	12.96	44.11
GPT-4o QD-IC + GPT-4o QA	<b>52.90</b>	9.85	37.25
CogVLM VQA	<b>64.18</b>	6.87	28.96
GPT-4o VQA	60.87	7.13	32.00
BLIP-2 VQA	47.03	7.31	45.67
GPT-4 VQA	49.24	10.43	40.15

### 3.5 Error Analysis

This section analyzes the inaccurate predictions of the best-performing IC-integrated pipeline, *GPT-4o QD-IC + GPT-4o QA*, to identify potential areas for improvement. Error analysis was conducted by evaluating the accuracy metric at the cosine similarity threshold of 0.7. The proportions of correct and incorrect predictions across different semantic and structural question category pairs are shown in the bar chart in Figure 13. According to the chart, the highest number of incorrect predictions, 2385, occurred in the *query-relation* category. This is followed by the *query-attribute* category with 1099 incorrect predictions and the *query-category* category pair with 509 incorrect predictions.

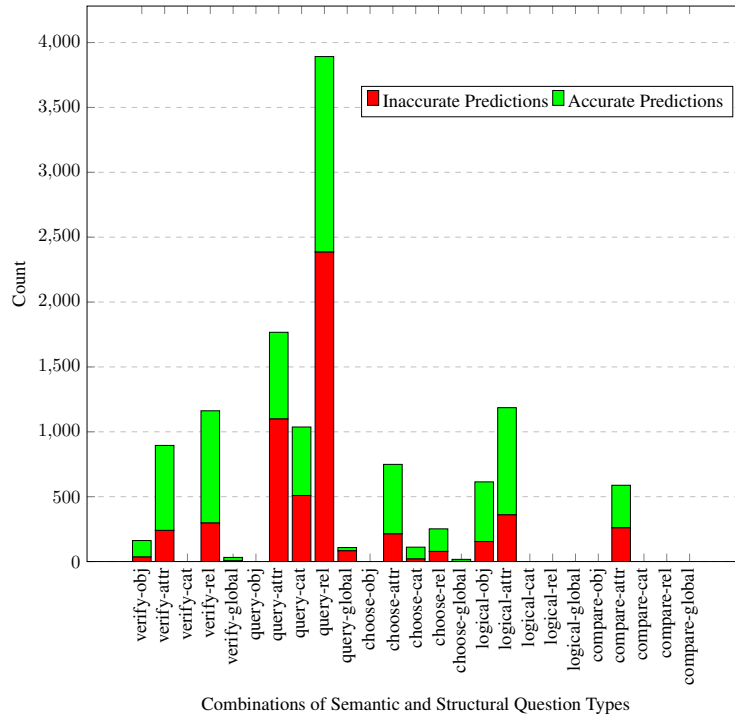


Figure 13: Distribution of incorrect and correct predictions of *GPT-4o QD-IC + GPT-4o QA* across question categories

100 incorrect predictions were randomly selected from the *query-rel*, *query-attr*, and *query-cat* question categories, and the reasons for these errors were analyzed.

The analysis revealed the following:

- 45 predictions were acceptable but did not match the label. See examples in Figure 14.
- 17 questions were answered as either "unknown" or "unclear". In 3 of these cases, the corresponding image caption contained the necessary details to answer the question. However, in the remaining 14 cases, the image captions did not include the relevant details. Among these, 2 questions involved details that were not present in the images. See examples in Figure 15.
- 29 predictions were incorrect because the associated image captions did not include the relevant information.
- 5 predictions were incorrect even though the image captions included the necessary details.
- 4 labels that did not match the predictions were detected as incorrect.

The high proportion of acceptable answers was due to the fact that the *GQA test-dev-balanced* dataset used for evaluation associates each question with only one correct answer, even though many questions can have multiple correct answers.

A significant portion of the errors stemmed from the image captions not being sufficiently informative. Although QD-IC provides general input about the context, it can be weak in fully conveying the

specific details asked by the questions. It was observed that the incorrectly answered questions were often related to less noticeable details in the scene.

"Unknown" and "unclear" responses also accounted for a significant portion of the errors. As noted in Section 3.4.4, these responses were more frequently given by GPT-4o compared to GPT-3.5. One reason for this was that the captions did not contain enough information to answer the questions adequately. To reduce the number of such "unknown" responses, it was necessary to improve the quality and informativeness of the image captions. Additionally, another type of "unknown" response occurred because GPT-4o tended to provide more specific answers; when it could not provide a precise answer, it responded with "unknown" instead of the expected general answer. For instance, when asked "*How big?*", the expected response might have been a general description such as "small" or "large." However, GPT-4o often responded with "unknown" if it was unable to provide an exact measurement.

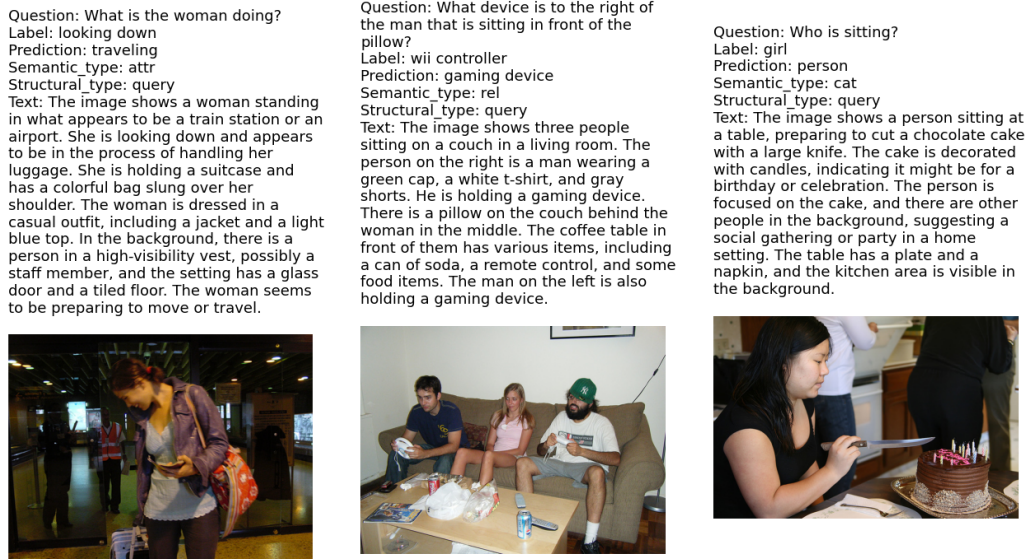


Figure 14: Examples of acceptable predictions that do not match the labels

### 3.6 Experiments with Different Prompt Designs

This section focuses on the investigation of prompt design aimed at enhancing the performance of MLLMs in the standard VQA task. Specifically, the VQA performance of a state-of-the-art MLLM, GPT-4o, will be assessed across various question categories with an error analysis conducted to identify potential improvement gaps. Based on the findings from the error analysis, prompts will be modified and their impacts on performance will be evaluated.

#### 3.6.1 Error Analysis with GPT-4o

This section takes a closer look at the incorrect predictions of GPT-4o on the GQA *test-dev-balanced* subset. Figure 16 compares the proportions of correct and incorrect predictions across semantic and

Question: What is the Asian person near the water bottle wearing?  
 Label: dress shirt  
 Prediction: unclear  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a young girl intently focused on a smartphone she is holding. She is wearing a black hoodie and has a bow in her hair. There are two adults beside her, one on each side. The adult on the left is wearing glasses and a striped shirt, and is gently touching the girl's shoulder. The adult on the right is also wearing a striped shirt and holding a camera. The setting appears to be indoors, possibly in a casual or home environment. There is no visible bottle or water in the image.



Question: What's the pipe made of?  
 Label: plastic  
 Prediction: unknown  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image shows a bathroom with a toilet and a sink vanity. The floor is covered with marble or marble-like tiles. The vanity is black with a white countertop and has a visible pipe underneath. The toilet is positioned next to a bathtub. The bathroom door is partially open, and the door has a brass or gold-colored hinge. The overall design is modern and clean.



Question: What's the skateboarder jumping off of?  
 Label: pavement  
 Prediction: none  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image depicts a vintage train car, labeled "FDR'S PERSONAL CAR," parked in front of a building with a large, illuminated sign that reads "ORIENT EXPRESS." The scene is set at night, with the area illuminated by streetlights and the sign. The train car is green with a silver roof, and the building behind it has a sign that also mentions "SIDETRACK ROOM." There is no skateboarder or any action of jumping visible in the image.



Figure 15: Examples of predictions containing "unknown" or "unclear"

structural question categories. The chart illustrates that incorrect predictions are proportionally higher in the *query-relation* (2097 instances), *query-attribute* (700 instances), and *query-category* (447 instances) question category pairs compared to others.

A set of 100 examples from the categories where GPT-4o made the most incorrect predictions was selected for error analysis. The following findings were obtained from this analysis:

- 52 answers were found to be acceptable even though they did not match the ground truth answers.
- 34 questions were incorrectly answered.
- 5 answers were predicted as "unknown", "unclear", or "I don't know."
- 7 questions had incorrect labels, for which GPT-4o mostly provided correct answers.
- 2 questions pertained to unclear elements in the image.

The analysis revealed that a significant proportion of the predictions contained acceptable answers despite not matching the ground-truth answer, as illustrated in Figure 18. As discussed in Section 3.5, this discrepancy arises because questions in the dataset can have multiple correct answers, even though only one is labeled as correct. As shown in Figure 19, some of these acceptable answers are even more detailed and accurate, further highlighting the limitations of the GQA test set and the evaluation method.



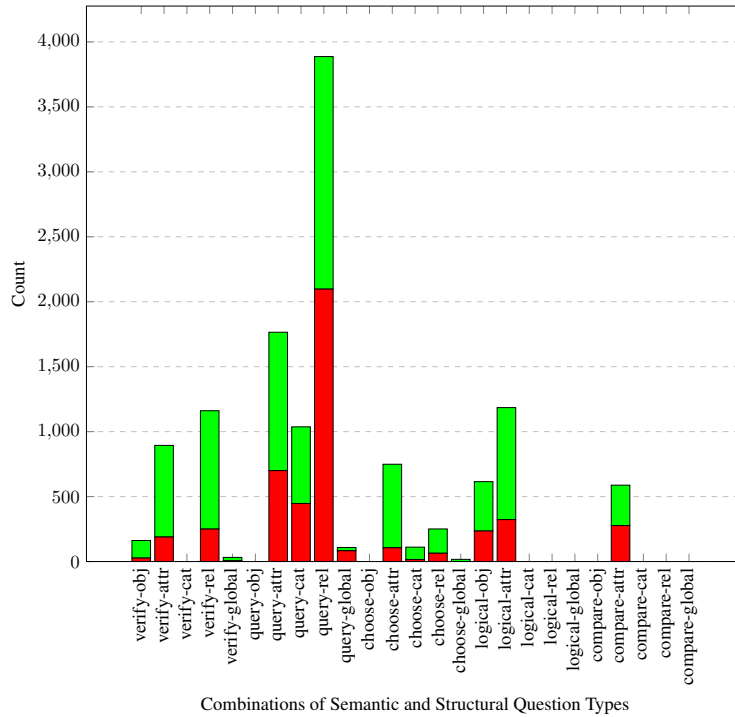


Figure 16: Incorrect predictions vs correct predictions of GPT-4o Vision across question combinations

A total of 417 instances were observed to be answered as "unknown", "unclear" or "I don't know", with 308 of these instances associated with the *query-relation* and *query-attribute* categories. A closer look reveals that the corresponding questions can be easily answered by a human in most cases. GPT-4o's inability to answer these questions while being able to answer more complex ones suggests that the model perceives the task as more complex than it is. As shown in the examples in Figure 17, GPT-4o may infer that a more specific answer is expected for the question "Who is holding the racket?", such as the name of the tennis player. In another example in Figure 17, GPT-4o may infer that more precise answers (similar to a measurement value in meters or a specific age) to the questions "What is the length of the train?" and "How old is the happy man?" are expected. Similar issues were encountered with GPT-4o in the QA phase of IC-integrated VQA, as mentioned in Section 3.5.

### 3.6.2 Designing GPT-4o Prompts for VQA

In this section, the impact of adding preliminary preparation steps to the prompt on GPT-4o's VQA performance will be evaluated based on the findings from the error analysis.

#### 3.6.2.1 Adding Object Identification and Hints to the Prompt

Based on the findings from the error analysis, the prompt was optimized to increase the performance of GPT-4o in VQA. Therefore, the prompt given in Figure 20 was designed by applying the following steps.

Question: Who is holding the racket?  
 Label: girl  
 Prediction: i don't know  
 Semantic\_type: rel  
 Structural\_type: query



Question: What is the length of the train tracks?  
 Label: long  
 Prediction: unknown  
 Semantic\_type: attr  
 Structural\_type: query



Question: How old is the happy man?  
 Label: old  
 Prediction: unknown  
 Semantic\_type: attr  
 Structural\_type: query



Figure 17: Examples of predictions containing "I don't know" or "Unknown"

Question: Where is the kid?  
 Label: street  
 Prediction: on the road  
 Semantic\_type: rel  
 Structural\_type: query



Question: What's the table in front of?  
 Label: chair  
 Prediction: people  
 Semantic\_type: rel  
 Structural\_type: query



Question: Which kind of furniture is not black?  
 Label: bed  
 Prediction: chair  
 Semantic\_type: cat  
 Structural\_type: query



Figure 18: Examples of acceptable predictions that did not match the labels

- The instruction to carefully read the question and identify objects and attributes relevant to the question was added to minimize the incorrect predictions in *attribute*, *category*, and *relation* categories. This guidance aims to help prevent GPT-4o from generating overly specific answers by restricting its response options to the objects and attributes present in the scene.
- Instructions were divided into steps, including reading the question, identifying the objects and attributes, and then answering the question. This step-by-step approach was designed to assist the model's reasoning process as a straightforward application of Chain of Thought (CoT) prompting.
- A frequently encountered issue is that GPT-4o often fails to provide an expected answer to relatively simple questions. As a result, it either gives an overly specific answer or responds

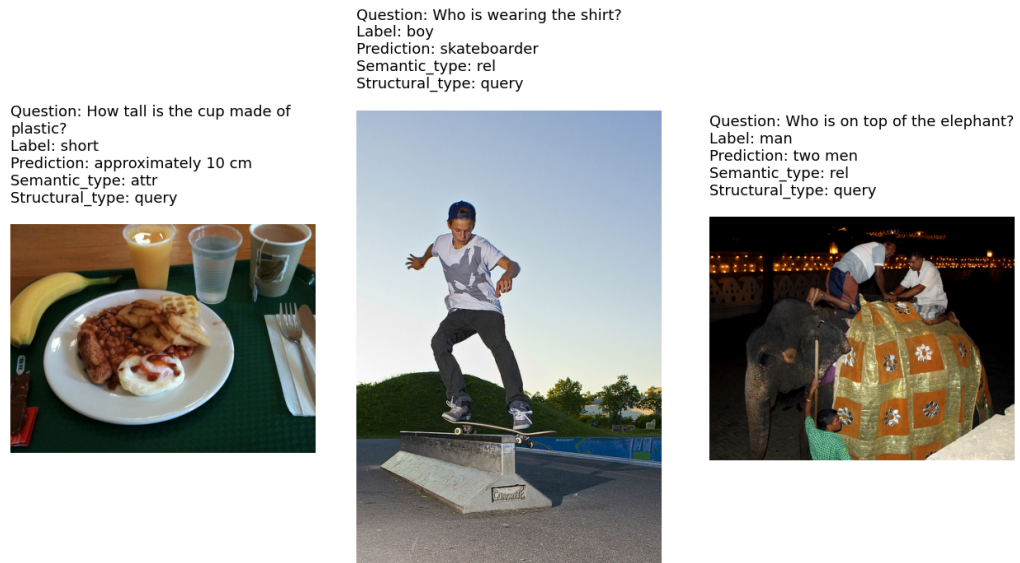


Figure 19: Examples of predictions giving more detailed and correct answers but did not match the labels

with "unknown" because it cannot detect sufficient specificity. Hints were provided to narrow down GPT-4o's response options and minimize "Unknown/I don't know" answers in order to obtain the expected response. Specifically, these hints include instructions for the MLLM to answer more broadly. For example, "How tall" questions are expected to be answered with basic categories such as "small", "medium" or "large" rather than specific measurements. Another common mistake encountered was with "Who" questions, which the model often answered with "Unknown" because it looked for more specificity, such as a person's name. Thus, hints were provided suggesting that responses to "Who" questions could relate to the person's gender or profession.

### 3.6.2.2 Adding a Vocabulary Definition Instruction to the Prompt

In another experiment, the prompt was optimized to minimize the "unknown" predictions to encourage the model to identify and use a targeted vocabulary of adjectives, verbs, nouns, and adverbs that describe objects, attributes, and relationships in the scene. This aims to generate answers that are contextually appropriate to the image content. Additionally, the instruction *"Make sure to provide broad and general responses rather than specific ones."* was added to encourage GPT-4o to avoid seeking specific details, such as exact measurements or person identification, and provide broader responses. The optimized prompt is shown in Figure 21.

<b>Object-Attribute Identification and Hint Provision in Prompt</b>
<p><b>System:</b> You are a helpful assistant providing concise 1-2 words answers based on the image content.  First, read the question.  Second, identify relevant objects and attributes in the scene related to the question.  Then, answer the question using objects and attributes identified.</p> <p>Hint: You don't need to give very detailed answers to the questions. For example, if the question asks how big an object is, it's enough to answer with 'small', 'medium' or 'large'. If the question asks 'who', the answer might be related to the 'gender' or 'occupation' of the person in the image. For instance, the answer to the question 'who is standing?' could be 'man' or the answer to 'who is wearing blue shorts?' could be 'player'.</p> <p><b>User:</b> [Question, Image]</p>

Figure 20: Adding object-attribute identification step and hints to the prompt

<b>Vocabulary Identification in Prompt</b>
<p><b>System:</b> You are a helpful assistant providing a concise answer to the question based on the image content.  <b>IMPORTANT:</b> Make sure that the answer is no more than two words.  <b>IMPORTANT:</b> Make sure that the answer is broad and general, not specific and detailed.  <b>IMPORTANT:</b> When generating an answer, your vocabulary must consist of adjectives, verbs, nouns, and adverbs that describe the objects, attributes, and relationships in the given image.</p> <p><b>User:</b> [Question, Image]</p>

Figure 21: Adding vocabulary definition step to the prompt

### 3.6.3 Results of Prompt Design Experiments

Table 8 presents the results obtained with different prompt designs. In the table, *P1* refers to the standard VQA prompt demonstrated in Figure 8. *P2* and *P3* refer to the prompts shown in Figure 20 and Figure 21, respectively. The best performance overall and in most question categories was obtained with the prompt *P2* containing object identification step and hints limiting the answer set. The highest performance improvement was observed in the *query* structural question category, with an increase of 4.29%. This improvement was followed by a 3.84 increase in the *relation*, a 3.18 increase in the *global*, and a 3.13% increase in the *category* questions. The main reason for this increase is that GPT-4o, with the instructions added to the prompt, generates responses that are more general and align with the elements of the scene. For example, GPT-4o initially responded to the question "Which aircraft is heavy?" with the answer "Chinook," referring to the model of the aircraft, instead of the correct answer, "helicopter." However, following the implementation of the *P2* and *P3* prompt designs, GPT-4o provided the correct answer.

The increase in the accuracy of predictions in the *global* questions, particularly *query-global* structural-semantic category, with the addition of the vocabulary definition stage in *P3* is found to be noteworthy. Dictionary definitions and instructions aimed at generating more general answers are helpful in this

category. For example, in a photo depicting a state police officer riding a motorcycle on a street, GPT-4o responds to the question "Which place is it?" with 'Mississippi,' based on the state police badge. In the global category, particularly for the frequently asked question "Which place is it?", GPT-4o with the standard prompt cannot produce general answers and often relies on scene details and Optical Character Recognition (OCR), frequently providing city and region names as responses. With the modified prompt, more general answers related to the location, such as "park", "street" or "sidewalk" can be obtained.

While the total number of "Unknown/I don't know" predictions was 417 with *P1*, *P3* containing the vocabulary identification step helped reduce this number to 351. *P2* had the most significant impact, reducing this number to 125. The results indicate that including instructions about context and vocabulary is helpful, but providing hints is more effective in guiding the model to produce answers in the expected answer set.

Table 8: Comparison of the performance of GPT-4o with different prompts

Question type	GPT-4o VQA P1	GPT-4o VQA P2	GPT-4o VQA P3
verify	78.91	<b>78.95</b>	77.93
query	50.99	<b>55.28</b>	53.28
choose	<b>83.51</b>	83.33	81.82
logical	68.89	68.05	<b>69.49</b>
compare	52.80	<b>54.83</b>	54.15
object	66.20	65.55	<b>66.32</b>
attribute	69.16	<b>70.08</b>	68.94
category	59.62	<b>62.75</b>	60.31
relation	54.41	<b>58.25</b>	56.63
global	43.31	46.49	<b>54.77</b>
total	61.58	<b>63.85</b>	62.63

### 3.7 Limitations

The evaluation in this thesis is conducted using the GQA dataset. While it is a comprehensive dataset, it may not cover all possible question types and scenarios in VQA. This could limit the generalizability of the findings to other VQA datasets. Additionally, GQA contains programmatically created questions, leading to vague question-image pairs and multiple correct answers. These ambiguities can impact the evaluation of the models and restrict the exploration of the full potential of LLMs/MLLMs in VQA.

Furthermore, this study utilizes pre-trained models such as CogVLM and BLIP-2, along with models like GPT-4o that are not publicly accessible. These models, which contain billions of parameters, have been trained on extensive datasets and may have memorized patterns within commonly used benchmark datasets like GQA. However, due to restricted access to the training data of many of these models, a thorough investigation of these potential issues was not feasible. This limitation could potentially affect the fairness of performance comparisons.

### 3.8 Summary and Discussion

This study proposes a new approach to VQA, referred to as question-driven image captioning-integrated (QD-IC-integrated) VQA. In this framework, context-aware image captions, which replace the images, are provided to the LLM as prompts. The study evaluates the performance of the QD-IC-integrated VQA approach across various types of questions that often require multi-step reasoning. The errors produced by the proposed method were analyzed to identify its limitations and potential gaps for further improvement. The performance of the proposed approach was compared with the standard VQA performance of MLLMs based on various evaluation metrics including LLM-based evaluation. Additionally, the VQA performance of GPT-4o, a recently developed state-of-the-art MLLM, was assessed. Errors generated by the model were analyzed, and prompts were optimized accordingly. The key findings from the experiments, addressing multiple aspects, are summarized in the following paragraphs.

**Overall performance:** The analysis conducted in this study demonstrates that question-driven image captioning offers a more contextually relevant representation of visual content in textual form compared to general-purpose image captioning. The effectiveness of this approach on overall performance highly depends on the accuracy and relevance of the image captions in representing the images to ensure precise answers to the questions. Although the proposed approach does not surpass the overall VQA performance of the state-of-the-art MLLMs, it remains competitive and can serve as an effective intermediary between images and LLMs. Overall performance can be further enhanced by customizing the image captioning models to more effectively process and utilize the contextual information provided by the questions.

**Performance across different question categories:** The proposed approach may offer advantages in providing answers to certain question categories. Experiments with GPT-4o, a state-of-the-art MLLM that has demonstrated superior performance in benchmarks, have shown that integrating image captions into the VQA enhances performance in specific question categories. Notably, it improves responses in the *object* category, which involves identifying objects within an image, as well as in the *logical* and *compare* categories. The positive impact is more pronounced in questions with a *logical-object* and *compare-attribute* structure. In these questions, the task typically involves determining which of two objects or attributes satisfies a given characteristic. At this point, context-aware image captions are useful in making the subjects of the question more explicit through text, thereby facilitating the comparison process. Performance improvement has also been found to be associated with questions that require details about an object that is not the main focus of the image. In such cases, the QD-IC approach highlights key terms, generating captions that make the relevant visual details more visible and help provide sufficient information for the QA phase. On the other hand, the analysis found that common errors are due to insufficient details in the image captions rather than the QA phase. While QD-IC provides general context, it might fail to fully address the specifics of the question.

Additionally, the performance of QD-IC depends on how informative the keywords extracted from the question are. For instance, in the *global* category, which predominantly comprises general questions such as "Which place is it?", the questions do not often include informative keywords. Conversely, categories like *object* and *logical* contain more keywords that are directly related to the question being asked. Future research could investigate the impact of using bigrams and trigrams as keywords, instead of unigrams, on the overall performance.

**Answer plausibility:** Analysis of both IC-integrated and standard VQA pipelines revealed that a significant portion of predictions made by LLMs/MLLMs were considered acceptable, even when they did not match the ground-truth labels. This discrepancy highlights deficiencies within the evaluation dataset. The study concludes that the development of more comprehensive and higher-quality datasets, along with refined evaluation metrics, is essential for accurately assessing the VQA performance of LLMs/MLLMs.

**Interpretability:** MLLMs exhibit high performance across various tasks, including VQA. However, interpreting the decision-making processes of these models, which have millions or even billions of parameters, poses a significant challenge. One effective strategy to enhance interpretability involves generating intermediate outputs, such as rationales, which can help elucidate the model's reasoning process and pinpoint sources of potential errors. In this context, using context-aware image captions as rationales facilitates interpretability and error analysis. This approach could be further enhanced in future research by integrating it with methods akin to Chain-of-Thought (CoT) prompting. Alternatively, interpretability could be improved by employing MLLMs exclusively for the image captioning stage while utilizing models with more transparent structures, such as graph neural networks, for the question-answering phase.

**Effect of prompt design:** Analysis indicates that MLLMs exhibit a significant dependence on prompt design in VQA tasks. Given that MLLMs engage in open-ended generation, incorporating hints to narrow down the output set, providing specific instructions for object identification and vocabulary definitions, and segmenting the VQA process into distinct steps to enable the LLM to think multiple times were found to have a positive impact on results. Future research could explore better prompt designs that help MLLMs better understand the expected output set and generate responses according to the desired context.





## CHAPTER 4

### CONCLUSION AND FUTURE WORK

This study investigates the efficacy of using image captions as an intermediary between images and LLMs in the VQA problem. In this context, the study proposes a question-driven image captioning approach that extracts keywords from the question and feeds them into a vision-language model to transfer the context of the question to the image captions.

Overall, although the performance of the proposed approach is lower than the standard VQA performance, it is competitive and has led to improvements in certain question categories. In future studies, leveraging open-source LLMs alongside question-driven image captions to analyze and visualize changes in attention weights can yield valuable insights. Furthermore, the caption generation process can be further improved to assist the LLM by providing a better, context-related rationale. Future work on caption generation for VQA may involve creating captions that encompass the chain-of-thought process required by the question.

In this thesis, experiments were conducted using LLMs and MLLMs with billions of parameters, which served as the foundation for performance analysis. Looking ahead, it would be beneficial to investigate the performance differences between standard VQA and captioning-integrated VQA in smaller-sized LLMs. Additionally, conducting prompt engineering studies to elucidate reasoning processes in smaller-sized MLLMs could provide valuable insights for future research.

Using image captions as an intermediary can offer various advantages beyond achieving better performance in certain question categories. Image captions serve as an interpretable linkage layer in the VQA pipeline, which facilitates error analysis during the QA phase. Additionally, separating VQA into two distinct modules -image captioning and question-answering- can provide flexibility. Combining text-based LLMs and vision-language models leverages the strengths of each and makes it easier to update modules with improved versions in the future.

On the other hand, within an IC-integrated VQA pipeline, the image captioning process continues to lack interpretability. Furthermore, the capacity of the pre-trained vision-language model used in image captioning is the most significant determinant of the pipeline’s overall performance. When a high-performing LLM is incorporated into this pipeline alongside a weak image captioning model, it demonstrates suboptimal performance because the intermediary outputs lack sufficient detail.

Image captioning and visual question answering are two challenging multimodal problems. Although they are closely related tasks, they each have different dynamics. Today, MLLMs are capable of performing these two tasks, as well as many other sub-tasks like visual grounding, in zero-shot settings. However, each of these tasks requires the MLLM to make task-specific inferences and form different

attention processes. When examining the IC-integrated VQA approach, it was found that most of the incorrect answers were due to deficiencies in image captioning rather than in the question-answering phase. The deficiencies in image captions were particularly identified in cases where the question was related to minor details in the image rather than the main focus/subject of the image. This indicates that current captioning models fail to fully grasp the important parts of the scene related to the question and need improvement.

A significant observation emerging from this study is the necessity for higher quality and more comprehensive datasets to assess the VQA performance of LLMs/MLLMs. Although the GQA dataset employed in this study contains semantically and structurally diverse questions that require reasoning skills, it may be inadequate for evaluating LLMs that generate open-ended responses. This inadequacy stems from its reliance on generalized answers (*e.g.*, expecting "large," "small," or "medium" as answers to questions about the length of an object) and the prevalence of questions that can have multiple correct answers. Therefore, a future research topic can be to develop evaluation sets to better assess the performance of LLMs/MLLMs in multimodal tasks encompassing VQA. Such evaluation could consider not only the accuracy of the predictions but also the plausibility of the intermediate outputs generated by the LLM as part of the reasoning process. Additionally, employing LLMs for evaluation and scoring in cases where questions have multiple correct answers could also be an avenue for future research.

Currently, as MLLMs scale, their performance and generalization capabilities are rapidly improving. However, understanding the decision-making mechanisms of these models is becoming increasingly difficult. As the competencies of MLLMs grow, their applications and commercial value also increase, leading large technology companies with high resource capacities to keep this technology proprietary, as seen with OpenAI's GPT-4. Nevertheless, the evaluation of MLLMs can only be done using inputs and outputs, underscoring the importance of evaluation sets. Designing evaluations based on inputs and outputs in a way that provides insights into the model's decision-making processes is crucial for gaining understanding, identifying the model's limitations, and pinpointing its weaknesses.

As the capacity of MLLMs continues to expand, their performance on benchmarks involving both NLP and multimodal tasks also increases. Currently, challenges persist in zero-shot tasks that involve multiple modalities and require multi-step reasoning. However, in the future, as the size of pre-training data and the number of model parameters increase, the issues seen in current benchmarks, including VQA, might be resolved. This could lead to the need for new, multi-task benchmarks to further explore the advanced capabilities and future frontiers of MLLMs.

## REFERENCES

- [1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [2] S. Amizadeh, H. Palangi, A. Polozov, Y. Huang, and K. Koishida, “Neuro-symbolic visual reasoning: Disentangling,” in *International Conference on Machine Learning*, pp. 279–290, Pmlr, 2020.
- [3] D. A. Hudson and C. 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*, pp. 6700–6709, 2019.
- [4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- [5] Y. Wang, M. Yasunaga, H. Ren, S. Wada, and J. Leskovec, “Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21582–21592, 2023.
- [6] W. Liang, Y. Jiang, and Z. Liu, “Graphvqa: Language-guided graph neural networks for scene graph question answering,” *NAACL-HLT 2021*, p. 79, 2021.
- [7] S. Barra, C. Bisogni, M. De Marsico, and S. Ricciardi, “Visual question answering: Which investigated applications?,” *Pattern Recognition Letters*, vol. 151, pp. 325–331, 2021.
- [8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat, *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [10] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, “Flamingo: a visual language model for few-shot learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
- [12] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, *et al.*, “Cogvlm: Visual expert for pretrained language models,” *arXiv preprint arXiv:2311.03079*, 2023.

- [13] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023.
- [14] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, 2024.
- [15] J. Huang and K. C.-C. Chang, “Towards reasoning in large language models: A survey,” *arXiv preprint arXiv:2212.10403*, 2022.
- [16] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang, “Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning,” *arXiv preprint arXiv:2401.06805*, 2024.
- [17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.
- [18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.
- [19] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, “Multimodal chain-of-thought reasoning in language models,” *arXiv preprint arXiv:2302.00923*, 2023.
- [20] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang, “Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 5168–5191, 2023.
- [21] S. Manmadhan and B. C. Kooor, “Visual question answering: a state-of-the-art review,” *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5705–5745, 2020.
- [22] H. Sharma and A. S. Jalal, “A survey of methods, datasets and evaluation metrics for visual question answering,” *Image and Vision Computing*, vol. 116, p. 104327, 2021.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [25] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.
- [26] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (A. Moschitti, B. Pang, and W. Daelemans, eds.), (Doha, Qatar), pp. 1724–1734, Association for Computational Linguistics, Oct. 2014.
- [27] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, “Are you talking to a machine? dataset and methods for multilingual image question,” *Advances in neural information processing systems*, vol. 28, 2015.

- [28] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, “Multimodal residual learning for visual qa,” *Advances in neural information processing systems*, vol. 29, 2016.
- [29] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, “Mutan: Multimodal tucker fusion for visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2612–2620, 2017.
- [30] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
- [31] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, “A simple neural network module for relational reasoning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [34] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “Videobert: A joint model for video and language representation learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7464–7473, 2019.
- [35] H. Tan and M. 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.
- [36] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137, Springer, 2020.
- [37] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal learning with transformers: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113–12132, 2023.
- [38] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “Mdetr-modulated detection for end-to-end multi-modal understanding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.
- [39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229, Springer, 2020.

- [40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [41] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 39–48, 2016.
- [42] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, “Learning to reason: End-to-end module networks for visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, pp. 804–813, 2017.
- [43] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, “Neural-symbolic vqa: Disentangling reasoning from vision and language understanding,” *Advances in neural information processing systems*, vol. 31, 2018.
- [44] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, “The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision,” in *International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, 2019.
- [45] Z. Zhao, K. Samel, B. Chen, *et al.*, “Proto: Program-guided transformer for program-guided tasks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 17021–17036, 2021.
- [46] T. Gupta and A. Kembhavi, “Visual programming: Compositional visual reasoning without training,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.
- [47] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, “Learning conditioned graph structures for interpretable visual question answering,” *Advances in neural information processing systems*, vol. 31, 2018.
- [48] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, “Language-conditioned graph networks for relational reasoning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10294–10303, 2019.
- [49] *The Cambridge Handbook of Intelligence*. Cambridge Handbooks in Psychology, Cambridge University Press, 2011.
- [50] K. Xu, J. Li, M. Zhang, S. S. Du, K.-i. Kawarabayashi, and S. Jegelka, “What can neural networks reason about?,” *arXiv preprint arXiv:1905.13211*, 2019.
- [51] C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, and C. Wolf, “How transferable are reasoning patterns in vqa?,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4207–4216, 2021.
- [52] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, “Collective generation of natural image descriptions,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 359–368, 2012.
- [53] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on attention for image captioning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4634–4643, 2019.

- [54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, pp. 2048–2057, PMLR, 2015.
- [55] Y. Xu, L. Li, H. Xu, S. Huang, F. Huang, and J. Cai, “Image captioning in the transformer age,” *arXiv preprint arXiv:2204.07374*, 2022.
- [56] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image captioning: Transforming objects into words,” *Advances in neural information processing systems*, vol. 32, 2019.
- [57] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [58] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 684–699, 2018.
- [59] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10685–10694, 2019.
- [60] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.
- [61] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, “Retrieving and reading: A comprehensive survey on open-domain question answering,” *arXiv preprint arXiv:2101.00774*, 2021.
- [62] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv preprint arXiv:1606.05250*, 2016.
- [63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [64] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [65] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [66] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.
- [67] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. 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*, pp. 6904–6913, 2017.

- [68] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.
- [69] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International journal of computer vision*, vol. 123, pp. 32–73, 2017.
- [70] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, “Ok-vqa: A visual question answering benchmark requiring external knowledge,” in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.
- [71] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, “Towards vqa models that can read,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- [72] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
- [73] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.
- [74] OpenAI, “Hello gpt-4o,” 2024.
- [75] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, *et al.*, “Bloom: A 176b-parameter open-access multilingual language model,” 2023.
- [76] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
- [77] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, “A survey on multimodal large language models,” *arXiv preprint arXiv:2306.13549*, 2023.
- [78] M. Luo, S. K. Sapat, R. Tallman, Y. Zeng, M. Vancha, A. Sajja, and C. Baral, “‘just because you are right, doesn’t mean i am wrong’: Overcoming a bottleneck in development and evaluation of open-ended vqa tasks,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.
- [79] C. Fellbaum, “Wordnet: An electronic lexical database,” *MIT Press google schola*, vol. 2, pp. 678–686, 1998.
- [80] H. Liu and P. Singh, “Conceptnet—a practical commonsense reasoning tool-kit,” *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.
- [81] S. Chen, M. Jiang, J. Yang, and Q. Zhao, “Attention in reasoning: Dataset, analysis, and modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7310–7326, 2021.



- [82] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.
- [83] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. 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*, pp. 10867–10877, 2023.
- [84] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi, “Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training,” *arXiv preprint arXiv:2210.08773*, 2022.
- [85] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, “An empirical study of gpt-3 for few-shot knowledge-based vqa,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3081–3089, 2022.
- [86] N. Rotstein, D. Bensaïd, S. Brody, R. Ganz, and R. Kimmel, “Fusecap: Leveraging large language models for enriched fused image captions,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5689–5700, 2024.
- [87] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [88] M. Grootendorst, “Keybert: Minimal keyword extraction with bert,” *Zenodo*, 2020.
- [89] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [90] C. Wang, S. Cheng, Q. Guo, Y. Yue, B. Ding, Z. Xu, Y. Wang, X. Hu, Z. Zhang, and Y. Zhang, “Evaluating open-qa evaluation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [91] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” 2023.
- [92] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” *arXiv preprint arXiv:2310.03744*, 2023.
- [93] O. Mañas, B. Krojer, and A. Agrawal, “Improving automatic vqa evaluation using large language models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 4171–4179, 2024.



# APPENDIX A

## ADDITIONAL EXAMPLES OF INCORRECT AND CORRECT PREDICTIONS

### A.1 Predictions Obtained with the QD-IC Integrated VQA Approach

#### A.1.1 GPT-4o QD-IC + GPT-4o QA

Question: What animal is inside the zoo?  
 Label: elephant  
 Prediction: elephant  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows an elephant inside a zoo enclosure. The elephant is walking on a sandy surface, and there are green bushes and a wooden structure in the background. The enclosure is surrounded by a fence with vertical posts and horizontal bars. The elephant appears to be in a well-maintained and spacious area.



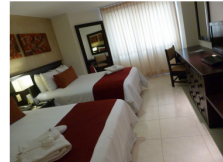
Question: What fruit is it?  
 Label: banana  
 Prediction: banana  
 Semantic\_type: cat  
 Structural\_type: query  
 Text: The image shows a meal served on a green tray. The meal includes a plate with baked beans, a poached egg, a sausage, some potato wedges, and a waffle. To the side, there is a whole banana, a cup of orange juice, a cup of water, and a cup of coffee. The tray also has a set of utensils (knife and fork) placed on a napkin. The banana is the fruit present in this meal.



Question: What is in front of the bench?  
 Label: fence  
 Prediction: fence  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a person playing softball, swinging a bat to hit a ball. The person is wearing a light blue shirt, blue shorts, black shoes, and sunglasses. They are standing on a dirt field, and there is a chain-link fence in the background. Behind the fence, there is a bench. The scene appears to be in a park or a sports field.



Question: What device is on the right of the curtains?  
 Label: television  
 Prediction: television  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image depicts a hotel room with two neatly made beds, each with white linens and a red throw blanket. The room features a large window with light-colored curtains, allowing natural light to filter through. To the right of the image, there is a dark wooden desk with a flat-screen device, likely a television, placed on it. The room also includes a mirror, a chair, and a bedside table with a lamp. The overall decor is modern and clean, with a neutral color palette and some artwork on the walls.



Question: Are there people behind the tennis rackets?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image shows three people lying on a tennis court, each with a tennis racket in front of them. They are all smiling and appear to be enjoying themselves. The person on the left is wearing an orange shirt, the person in the middle is wearing sunglasses and a white headband, and the person on the right is wearing a pink shirt and a headband with the word "HEAD" on it. The background features trees and a clear sky, indicating a sunny day.



Question: Are there glasses or women?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: obj  
 Structural\_type: logical  
 Text: The image shows a young baseball player in a green jersey and a black helmet swinging a bat. In the background, there are a few spectators, including women wearing glasses, sitting on chairs and watching the game. The scene appears to be outdoors, likely at a baseball field.



Question: Is the backpack large and black?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: logical  
 Text: The image shows a person skiing down a snowy slope. The skier is wearing a red jacket, black pants, and a black helmet. They are also carrying a large black backpack. The skier is in motion, leaning forward and using ski poles for balance. In the background, there are other skiers and a forest of tall trees. The scene is bright and sunny, indicating a clear day on the ski slope.



Question: Does the jersey look long sleeved and blue?  
 Label: no  
 Prediction: no  
 Semantic\_type: attr  
 Structural\_type: logical  
 Text: The image shows a baseball player in mid-air, seemingly in the process of throwing a baseball. He is wearing a blue jersey with the word "San Diego" written on it, white pants, blue socks, and a blue cap. The jersey is short-sleeved, not long-sleeved. The player is on a grassy field with a dirt path, likely a baseball field.



Figure 22: Examples of **correct** predictions by *GPT-4o QD-IC + GPT-4o QA*

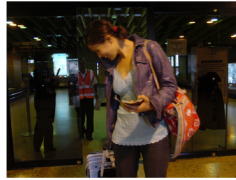
Question: What is the gender of the person in the middle?  
 Label: male  
 Prediction: unknown  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image shows a person performing a skateboard trick over a green trash can on a sidewalk. The person is in mid-air, with their skateboard tilted and their body positioned above the trash can. The person is wearing a black cap, a light-colored shirt, and dark pants. In the background, there are several parked cars, a large building, and some trees. The scene appears to be in an urban area with a mix of greenery and architecture.



Question: What is the cone in front of?  
 Label: machine  
 Prediction: sidewalk  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image depicts a person standing next to an orange fire hydrant on a sidewalk. The hydrant is releasing a stream of water, which the person is interacting with. The person is dressed in a sleeveless top and shorts. The scene is set in an urban area with a construction site in the background, which is fenced off. There are several orange traffic cones placed along the sidewalk and street, likely to indicate construction or maintenance work. The background also includes construction equipment and a tree next to the sidewalk.



Question: Who is wearing the jacket?  
 Label: woman  
 Prediction: person  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a person wearing a purple jacket. They are standing indoors, possibly at a train station or an airport, as they are holding a suitcase and a bag. The person appears to be looking down at something they are holding in their hands. The background includes a person in a high-visibility vest and a glass door.



Question: Does the shirt look yellow?  
 Label: no  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: verify  
 Text: The image shows a busy urban street scene with several buses parked along the side of the road. In the foreground, two people are crossing the street. One of them is wearing a yellow shirt. The buses are of different colors, with one being a light blue Volvo and another being a beige and green bus. The background features a building with large windows and some trees. The overall scene appears to be a typical city environment with public transportation and pedestrians.



Question: Who is in front of the tree?  
 Label: man  
 Prediction: unknown  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a Siamese cat standing next to a pair of black and white sneakers. The cat is looking directly at the camera. In the background, there is a person whose legs and arms are partially visible, and some greenery, including a fern, is present. The scene appears to be outdoors, possibly on a patio or a porch. The context keyword "tree" is not directly visible in the image.



Question: What is the skateboarder skating on?  
 Label: skateboard  
 Prediction: ramp  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image shows a skate park during its grand opening event. Several skateboarders are present, with one prominently holding a skateboard while standing on a ramp. The ramp is designed with a smooth, curved surface, ideal for performing tricks. The park is outdoors, with trees and a fence visible in the background. The sign on the ramp emphasizes the grand opening and includes a reminder that helmets are required for safety. The atmosphere appears lively, with multiple people gathered, likely excited to try out the new skate park.



Question: What's lying next to the building?  
 Label: road  
 Prediction: unknown  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image depicts a red and yellow Crosville bus parked on a paved area. The bus is labeled with the route number 16 and the destination "Saltney." The bus is a Plaxton Pointer model with the registration number P975 MBF. In the background, there is a large brick building with a window and some architectural details. The bus is stationary, and the area appears to be a bus stop or a designated parking area.



Question: In which part is the open umbrella?  
 Label: left  
 Prediction: none  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image shows a person sitting at a table, eating a sandwich or burger. The table has a green basket with a paper liner, a bottle of Cheerwine soda, and two condiment bottles, one with yellow mustard and the other with red ketchup. The person is wearing a gray T-shirt and appears to be enjoying their meal. The setting seems to be a casual dining area with a rustic or cozy atmosphere. There are no visible umbrellas or open umbrellas in the image.



Question: Is the spoon on top of the food below the grapes?  
 Label: yes  
 Prediction: no  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: This image shows three people posing together and smiling. The context keywords "spoon", "grapes", and "food" do not appear to be directly related to the image. The individuals are dressed formally, with one person wearing a graduation gown and the others in formal attire. The background is a plain, neutral-colored backdrop, suggesting the photo might have been taken at a formal event or celebration.



Question: Is the tall tree on the right?  
 Label: no  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: verify  
 Text: The image shows a yellow pedestrian crossing sign that is slightly tilted. The sign is attached to a tall tree on the right side of the image. In the background, there are several tall palm trees against a partly cloudy sky. The scene appears to be in an urban area with some buildings visible in the lower part of the image.



Question: Does the blue sky look cloudless?  
 Label: no  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: verify  
 Text: The image depicts an urban scene with a river running through the city. There are multiple trains visible, one on a bridge crossing the river and another on a parallel track. The cityscape includes a mix of modern and older buildings, with some high-rise structures in the background. The sky above the city is clear and cloudless, showcasing a bright blue color. The overall atmosphere is vibrant and bustling.



Question: Are there any horses in front of the white fence?  
 Label: no  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image depicts a harness racing event with several horses pulling sulkeys (two-wheeled carts) driven by jockeys. The horses are racing on a dirt track. In the background, there is a white fence that runs parallel to the track, and a white vehicle is parked near the fence. The sky is mostly clear with a few clouds, and there are trees and grassy areas beyond the fence.



Figure 23: Examples of incorrect predictions by GPT-4o QD-IC + GPT-4o QA

## A.1.2 CogVLM-C QD-IC + GPT-3.5 QA

Question: Are there either women or men that are eating?  
 Label: no  
 Prediction: no  
 Semantic\_type: obj  
 Structural\_type: logical  
 Text: The scene in the image depicts a woman standing next to a fire hydrant, seemingly interacting with or observing the water flowing from it. The environment appears urban with construction activity in the background, as indicated by the fencing and construction vehicles.



Question: What animal is wearing the hat?  
 Label: dog  
 Prediction: dog  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image showcases a white fluffy dog wearing a red Santa hat, sitting next to a decorated Christmas tree. The dog appears joyful, with its tongue out and eyes wide open. In the background, there's a snowman decoration and other festive items, suggesting a holiday setting.



Question: Is the shirt soft and white?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: logical  
 Text: The image showcases a Siamese cat with a soft, cream-colored coat and darker face markings. The cat is positioned next to a pair of worn-out shoes, possibly indicating that it has been playing or exploring the area. In the background, there's a person wearing a white shirt, and some greenery, suggesting an outdoor setting.



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: What does the chubby man hold?  
 Label: umbrella  
 Prediction: umbrella  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The scene in the image features two men, presumably monks, walking on a wet road. They are both holding umbrellas to shield themselves from the rain. The road appears to be lined with trees and there's a banner overhead with some text in a foreign language.



Question: Are the drawers underneath the countertop?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image showcases a kitchen with wooden cabinets, a granite countertop, and various kitchen appliances. The cabinets have multiple drawers, and there's a sink with a faucet underneath the countertop.



Question: Is the person to the right of the cup wearing jeans?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image depicts a person wearing jeans, sitting at a table with a cup in front of them. The person appears to be working on a laptop, with multiple beverages and a bottle nearby.



Question: Are there beds next to the small outlet?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image showcases a young child sitting on a green bed in a room. The child is wearing a black shirt with a graphic design and brown pants. In the background, there's a blue toy cabinet with a toy on it. An electrical outlet is visible on the wall next to the bed.



Question: How big are the pickles that are on top of the cutting board?  
 Label: small  
 Prediction: small size  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image showcases a meal setting on a wooden board. There's a burger with a bun topped with a skewer, a side of thick-cut fries placed in a metal basket, and a small serving of pickles and onions. The ambiance suggests a casual dining environment, possibly a restaurant or cafe.



Question: Which side is the mirror on?  
 Label: right  
 Prediction: right side  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image showcases a group of people gathered around a red motorcycle. One of the motorcycle's mirrors is prominently visible in the foreground. The individuals seem to be engaged in a conversation, and the motorcycle has a distinctive license plate.



Question: Are there both a horse and a fence in the image?  
 Label: yes  
 Prediction: yes  
 Semantic\_type: obj  
 Structural\_type: logical  
 Text: The image showcases a race track where horses are being driven in harnesses. There's a white fence in the background, and the sky is clear with a few clouds. The horses are in motion, and there are spectators or participants in the background.



Question: What appliance is the refrigerator larger than?  
 Label: stove  
 Prediction: oven  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image showcases a kitchen undergoing renovation. Key elements include a refrigerator, an oven, and exposed wooden beams. The space appears to be in the midst of construction or repair, with insulation materials and tools scattered around.



Figure 24: Examples of correct predictions by CogVLM-C QD-IC + GPT-3.5 QA

Question: Who is wearing a wristband?  
 Label: woman  
 Prediction: three individuals  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image showcases three individuals on a tennis court, lying down and posing with their hands on their chins. They are wearing sporty attire, including wristbands and sunglasses. The background features a clear sky and green trees, suggesting a sunny day at a park or tennis facility.



Question: What color is the dirt?  
 Label: red  
 Prediction: brown  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image showcases a baseball game in progress. The field is covered in dirt, which is typical for baseball diamonds. The colors in the image are vibrant, with the green of the grass, the brown of the dirt, and the white of the baseball uniforms.



Question: What is beneath the microwave?  
 Label: bananas  
 Prediction: wooden cabinet  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image showcases a kitchen setting with a microwave placed beneath a wooden cabinet. The cabinet has glass panels, and the floor is adorned with tiles. There's also a bookshelf filled with books in the background.



Question: What's the man doing?  
 Label: standing  
 Prediction: observing/waiting  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The scene in the image shows a group of people waiting at a traffic light. One man is prominently standing in the foreground, possibly observing or waiting for the traffic to move.



Question: Is the van in front of a balloon?  
 Label: no  
 Prediction: yes  
 Semantic\_type: rel  
 Structural\_type: verify  
 Text: The image showcases a vintage van, possibly a Volkswagen Type 2, parked outdoors. The van is adorned with various decorations, including a sign that reads 'SNOOP' and another that says 'Sweet Wheels'. There are two individuals standing in front of the van, seemingly interacting with it. In the background, there are tents and other structures, suggesting this might be a market or festival setting. Additionally, there's a balloon-like object on the left side of the image.



Question: Which kind of clothing is not pink?  
 Label: hat  
 Prediction: black shorts  
 Semantic\_type: cat  
 Structural\_type: query  
 Text: The image showcases a female tennis player in action on a tennis court. She is wearing a pink tank top and black shorts, and she is holding a green tennis racket. The player appears focused and is in the midst of a swing, suggesting she is playing a match or practicing.



Question: Is the bag made of leather lying on top of a sofa?  
 Label: no  
 Prediction: green leather  
 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 is inside the bowl to the right of the beans?  
 Label: cookies  
 Prediction: white sauce  
 Semantic\_type: rel  
 Structural\_type: query  
 Text: The image showcases a lunchbox with various food items neatly arranged. There are green beans placed inside a white container, a bowl of white sauce on the right, and other items like carrots, sushi rolls, and cupcakes on the left.



Question: Do you see any skis?  
 Label: yes  
 Prediction: no  
 Semantic\_type: obj  
 Structural\_type: verify  
 Text: The image showcases a seaplane docked at a waterfront, with a backdrop of residential buildings and trees. The seaplane has the label 'SYDNEYSKIPLANES.COM' and the registration 'VH-SVP'. The overall ambiance suggests a serene waterfront setting, possibly in a city like Sydney, given the label on the plane.



Question: Are the shorts large and blue?  
 Label: no  
 Prediction: blue shorts  
 Semantic\_type: attr  
 Structural\_type: logical  
 Text: The image showcases a woman standing on a sandy beach, holding a large surfboard with a red top and flower designs. She is wearing blue shorts and a red rash guard. Behind her, there's a large white banner with a blue logo. The ocean waves can be seen crashing in the background, and there are hills or mountains further away.



Question: Are there any red fire trucks?  
 Label: no  
 Prediction: yes  
 Semantic\_type: obj  
 Structural\_type: verify  
 Text: The image showcases a vibrant urban scene with a prominent red fire truck in the foreground. The truck has the number '65' on it and is labeled 'THE MID TOWN MOB'. Behind the truck, there are other vehicles, including a black van. The American flags can be seen fluttering in the background, suggesting a patriotic or national event. The scene is bustling with activity, and the sunlight creates a dramatic effect, casting shadows and illuminating the truck's details.



Question: On which side is the picture?  
 Label: left  
 Prediction: indoor kitchen  
 Semantic\_type: attr  
 Structural\_type: query  
 Text: The image showcases an indoor setting, possibly a kitchen, where a person is pouring a liquid from a bottle into a blender. There's a green balloon hanging above, and various kitchen items are visible in the background.



Figure 25: Examples of **incorrect** predictions by *CogVLM-C QD-IC + GPT-3.5 QA*

## A.2 Predictions Obtained with the standard VQA pipeline

### A.2.1 GPT-4o VQA





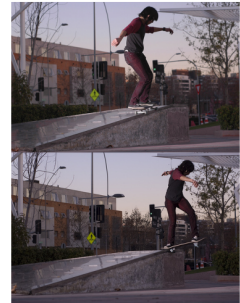







<p>Question: Is the jacket white and open?          Label: yes          Prediction: yes          Semantic_type: attr          Structural_type: logical</p> 	<p>Question: Is the carpet yellow and soft?          Label: no          Prediction: no          Semantic_type: attr          Structural_type: logical</p> 	<p>Question: Who is talking on the device that looks black and silver?          Label: child          Prediction: child          Semantic_type: rel          Structural_type: query</p> 	<p>Question: How clean is the sidewalk that the sign is in front of?          Label: dirty          Prediction: dirty          Semantic_type: attr          Structural_type: query</p> 
<p>Question: On which side of the picture is the traffic cone?          Label: right          Prediction: right          Semantic_type: attr          Structural_type: query</p> 	<p>Question: Which color is the mat, red or blue?          Label: red          Prediction: red          Semantic_type: attr          Structural_type: choose</p> 	<p>Question: Is she wearing a skirt?          Label: yes          Prediction: yes          Semantic_type: rel          Structural_type: verify</p> 	<p>Question: What aircraft is military?          Label: helicopter          Prediction: helicopter          Semantic_type: cat          Structural_type: query</p> 
<p>Question: The black chicken is standing on top of what?          Label: desk          Prediction: desk          Semantic_type: rel          Structural_type: query</p> 	<p>Question: What is the item of furniture that is below the folded towel near the sink?          Label: table          Prediction: table          Semantic_type: rel          Structural_type: query</p> 	<p>Question: Which kind of vehicle is old?          Label: van          Prediction: van          Semantic_type: cat          Structural_type: query</p> 	<p>Question: Is the canopy dark and dry?          Label: yes          Prediction: yes          Semantic_type: attr          Structural_type: logical</p> 

Figure 26: Examples of **correct** predictions by *GPT-4o VQA*

Question: Is the fruit on top of the tray yellow and thick?  
 Label: no  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: logical



Question: What is this fence in front of?  
 Label: bench  
 Prediction: softball field  
 Semantic\_type: rel  
 Structural\_type: query



Question: Is the flower pot on the shelf gray or orange?  
 Label: orange  
 Prediction: gray  
 Semantic\_type: attr  
 Structural\_type: choose



Question: Who is wearing shorts?  
 Label: woman  
 Prediction: person in a red  
 Semantic\_type: rel  
 Structural\_type: query



Question: What kind of furniture is to the right of the lamp above the table?  
 Label: bed  
 Prediction: chair  
 Semantic\_type: rel  
 Structural\_type: query



Question: What are the people behind the rackets lying on top of?  
 Label: courtyard  
 Prediction: tennis court  
 Semantic\_type: rel  
 Structural\_type: query



Question: Which kind of vehicle is parked behind the car?  
 Label: van  
 Prediction: motorcycle  
 Semantic\_type: rel  
 Structural\_type: query



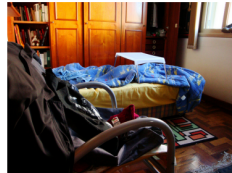
Question: Do you see a large mirror or table?  
 Label: no  
 Prediction: table  
 Semantic\_type: obj  
 Structural\_type: logical



Question: Does the soap seem to be white and large?  
 Label: no  
 Prediction: yes  
 Semantic\_type: attr  
 Structural\_type: logical



Question: What piece of furniture is not covered?  
 Label: bed  
 Prediction: table  
 Semantic\_type: cat  
 Structural\_type: query



Question: Which kind of furniture is to the right of the table?  
 Label: shelves  
 Prediction: cabinet  
 Semantic\_type: rel  
 Structural\_type: query



Question: What is the person walking on?  
 Label: ground  
 Prediction: sidewalk  
 Semantic\_type: rel  
 Structural\_type: query



Figure 27: Examples of **incorrect** predictions by *GPT-4o VQA*



## APPENDIX B

### ADDITIONAL RESULTS

#### B.1 Performance Evaluation Using Cosine Similarity Threshold of 0.8 and Exact Match Criteria

Table 9: Comparison of the accuracy values at cosine similarity threshold 0.8 for various IC-integrated VQA pipelines on GQA *test-dev-balanced*.

Question type	CogVLM-C GP-IC + GPT-3.5 QA	CogVLM-V GP-IC + GPT-3.5 QA	CogVLM-C <b>QD-IC</b> + GPT-3.5 QA	CogVLM-C <b>QD-IC</b> + GPT-4o QA	FuseCap GP-IC + GPT-3.5 QA	BLIP-2 GP-IC + GPT-3.5 QA	GPT-4 <b>QD-IC</b> + GPT-4o QA	GPT-4o <b>QD-IC</b> + GPT-4o QA	GPT-4o GP-IC + GPT-4o QA
verify	63.01	58.52	67.40	64.87	53.59	55.81	69.58	<b>74.33</b>	31.77
query	28.40	26.78	30.16	31.22	23.07	25.40	32.02	<b>35.79</b>	35.59
choose	60.54	57.80	61.34	57.26	54.07	56.82	58.77	<b>68.43</b>	52.92
logical	59.51	60.28	59.67	63.72	57.07	56.07	68.16	<b>71.32</b>	64.78
compare	51.27	51.44	50.93	52.12	53.65	47.70	53.82	<b>55.51</b>	48.38
object	61.95	63.23	59.38	68.63	59.38	58.35	71.20	<b>75.44</b>	72.23
attribute	46.44	44.12	49.29	46.55	41.77	41.83	52.26	<b>56.16</b>	44.55
category	47.35	40.20	41.86	45.86	36.90	39.16	44.29	<b>48.65</b>	47.78
relation	37.37	35.30	39.03	40.46	30.48	33.53	39.92	<b>44.64</b>	41.07
global	<b>46.49</b>	42.67	41.40	39.49	42.03	45.22	40.12	42.67	45.85
total	43.01	41.21	44.83	45.22	37.64	39.11	47.40	<b>51.67</b>	45.12

Table 10: Comparison of the EM accuracy values for various IC-integrated VQA pipelines on GQA *test-dev-balanced*.

Question type	CogVLM-C GP-IC + GPT-3.5 QA	CogVLM-V GP-IC + GPT-3.5 QA	CogVLM-C <b>QD-IC</b> + GPT-3.5 QA	CogVLM-C <b>QD-IC</b> + GPT-4o QA	FuseCap GP-IC + GPT-3.5 QA	BLIP-2 GP-IC + GPT-3.5 QA	GPT-4 <b>QD-IC</b> + GPT-4o QA	GPT-4o <b>QD-IC</b> + GPT-4o QA	GPT-4o GP-IC + GPT-4o QA
verify	62.87	58.52	67.18	64.87	53.60	55.81	69.58	<b>74.33</b>	64.87
query	17.72	18.00	18.60	25.87	15.42	18.07	25.37	<b>29.14</b>	26.62
choose	56.56	55.49	56.82	54.78	51.50	53.19	56.91	<b>66.66</b>	51.41
logical	59.23	60.23	59.01	63.72	57.01	55.96	68.16	<b>71.32</b>	64.78
compare	50.42	50.93	49.06	52.12	51.78	47.19	53.82	<b>55.51</b>	48.38
object	61.56	63.23	58.74	68.63	59.25	58.22	71.20	<b>75.44</b>	72.23
attribute	42.64	42.37	44.72	44.93	39.44	38.80	50.41	<b>58.10</b>	43.62
category	29.06	27.58	29.76	39.94	27.06	29.59	34.55	39.86	<b>40.81</b>
relation	29.14	28.05	30.27	35.98	24.51	28.42	34.96	<b>39.65</b>	36.68
global	34.39	38.21	30.57	40.12	33.75	38.21	40.12	40.76	<b>42.03</b>
total	36.77	36.21	37.94	42.10	33.17	34.77	43.63	<b>47.90</b>	42.20

## B.2 Comparison of GPT-4o QD-IC + GPT-4o QA with GPT-4o VQA in Question Categories

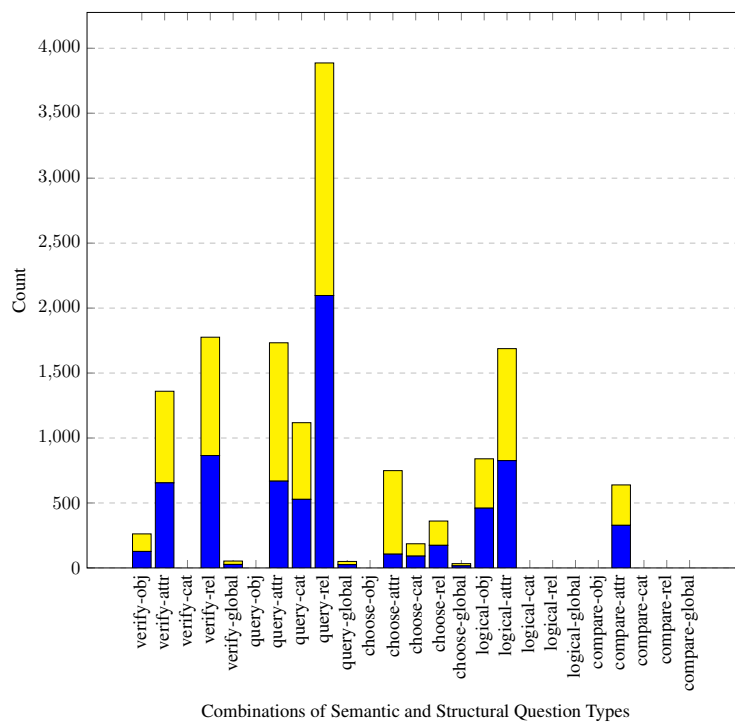


Figure 28: Number of correct predictions of GPT-4o QD-IC + GPT-4o QA (blue) vs GPT-4o VQA (yellow) across different structural-semantic question categories