OAD-Promoter: Enhancing Zero-Shot VQA Using Large Language Models with Object Attribute Description

Quanxing Xu[1], Ling Zhou[1*], Feifei Zhang[2], Jinyu Tian[1], Rubing Huang[1,3]
[1]School of Computer Science and Engineering, Macau University of Science and Technology, Macau SAR, China
[2]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China
[3]Macau University of Science and Technology Zhuhai MUST Science and Technology Research Institute, Zhuhai, China
lzhou@must.edu.mo

Abstract

Large Language Models (LLMs) have become a crucial tool in Visual Question Answering (VQA) for handling knowledge-intensive questions in few-shot or zero-shot scenarios. However, their reliance on massive training datasets often causes them to inherit language biases during the acquisition of knowledge. This limitation imposes two key constraints on existing methods: (1) LLM predictions become less reliable due to bias exploitation, and (2) despite strong knowledge reasoning capabilities, LLMs still struggle with out-of-distribution (OOD) generalization. To address these issues, we propose **O**bject **A**ttribute **D**escription **Promoter** (OAD-Promoter), a novel approach for enhancing LLM-based VQA by mitigating language bias and improving domain-shift robustness. OAD-Promoter comprises three components: the Object-concentrated Example Generation (OEG) module, the Memory Knowledge Assistance (MKA) module, and the OAD Prompt. The OEG module generates global captions and object-concentrated samples, jointly enhancing visual information input to the LLM and mitigating bias through complementary global and regional visual cues. The MKA module assists the LLM in handling OOD samples by retrieving relevant knowledge from stored examples to support questions from unseen domains. Finally, the OAD Prompt integrates the outputs of the preceding modules to optimize LLM inference. Experiments demonstrate that OAD-Promoter significantly improves the performance of LLM-based VQA methods in few-shot or zero-shot settings, achieving new state-of-the-art results.

Introduction

Different from other VL tasks (e.g., video captioning (Zhong et al. 2023; Yang et al. 2025)), Visual Question Answering (VQA) (Antol et al. 2015; Goyal et al. 2017), serves as one of the most important representatives of the multi-modal field, where a natural language question accompanies an image, has attracted considerable attention in recent years due to its requirement for generating accurate natural language responses.

Nevertheless, language bias remains a significant challenge in this field. For instance, the dominant answer for the question type "What color ... bananas?" in the training data is "yellow". Consequently, VQA models may exploit this simple correlation, relying on superficial cues rather than integrating visual information to capture underlying semantics and perform reasoning, which can lead to incorrect answers.

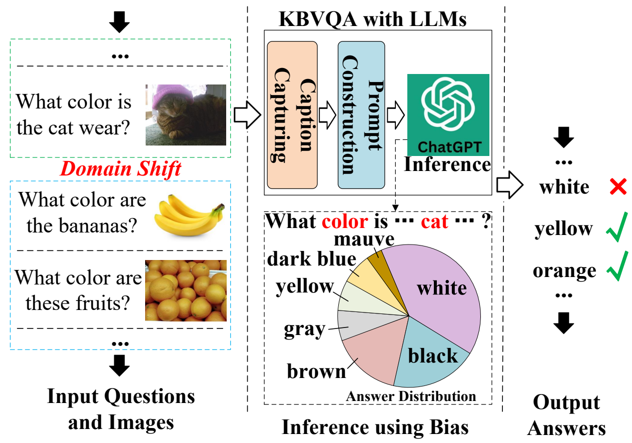

Figure 1: The illustration of the problem in existing LLM-based KBVQA. Like conventional VQA models, LLMs tend to exploit the inside language bias when they conduct inference. This drawback hampers both their accuracy and domain adaptation capabilities.

In addition, language bias is not only an essential issue in conventional VQA methods, but also a new emerging problem in Knowledge-based VQA (KBVQA) (Marino et al. 2019) methods using Large Language Models (LLMs). In recent years, LLM-based methods have attained notable advancements in this field. For instance, Yang et al. (Yang et al. 2022) proposed a few-shot pipeline for GPT-3 (Brown et al. 2020); Guo et al. (Guo et al. 2023) successfully realized a zero-shot VQA with OPT (Zhang et al. 2022). Despite these achievements being dazzling and impressive, few studies have acknowledged the bias problem underlying the pre-trained LLMs. Lately, with the rapid development of pre-trained large models (e.g., Multi-modal Large Language Models (MLLM), Vision-Language Pre-trained models (VLPs), Vision-Language Models (VLMs)), some issues about their drawbacks have come to the surface. The latest works (Zhang, Jiang, and Zhao 2024) have implied that the dataset-bias prob-

*Corresponding author.

lem can not be underestimated in studies with pre-trained large models. Since these large models were pre-trained on a vast scale of datasets and corpora, they inevitably acquire some spurious correlations instead of the target pattern during the training, which is called shortcut learning (Geirhos et al. 2020). Therefore, they naturally inherit the bias with the accumulation of learning. As illustrated in fig. 1, this drawback has two negative impacts on the LLM-based VQA methods. First, the prediction of LLMs is not reliable enough due to the exploitation of language bias. Second, although LLMs have demonstrated outstanding performance in knowledge reasoning, the out-of-distribution (OOD) circumstance remains a challenging issue, as language bias exacerbates the difficulty for LLMs in adapting to new domains.

The existing LLM-based KBVQA overlooked the combination of the global and regional visual information. Even though LLMs are equipped with tremendous knowledge, it is not easy for them to adapt to new domains seamlessly under a distribution-changing scenario. The invention of a subsidiary memory module to assist LLMs in domain-shift prediction has not been investigated in previous works. Inspired by the impressive work (Selvaraju et al. 2019; Wu and Mooney 2019; Li et al. 2024; Zhang, Zhang, and Xu 2023), we recognize that the multi-granularity captions can enhance the integrity and richness of visual information, thereby reducing the impact of language bias; making the memory examples available for the new inference is a refined knowledge supplement for LLMs, consequently improving the reliability of the prediction, especially in domain-shifting situations.

Motivated by the above factors, we posit that 1) More meticulous visual information can mitigate the language bias. 2) Memory examples assistance can benefit the inference and improve the prediction reliability. 3) A prompt that includes the above two characteristics can constantly promote domain-shift capacity with the inference proceeding. Accordingly, we propose **O**bject **A**ttribute **D**escription **Promoter** (OAD-Promoter), a LLM-based multiple-module collaborative zero-shot approach designed to overcome inherent language bias in LLMs via a multi-granularity visual description input; and to constantly improve the domain adaptive capability of LLMs via assistant examples support.

The goal of this work is realized by the combination of three components: the Object-concentrated Example Generation (OEG) module, the Memory Knowledge Assistance (MKA) module, and the OAD Prompt.

Note that the entire OAD-Promoter procedure does not rely on any external knowledge sources or data that need to be retrieved. It is a zero-shot approach.

Our key contributions can be summarized in fourfold:

- We introduce multi-granularity captions to LLM-based VQA and propose the OEG module, leveraging enhanced visual information to mitigate inherited language biases in LLMs.
- We design the MKA module to exploit relevant stored object-attribute examples, improving prediction reliability through memory-augmented knowledge support.
- We develop the OAD Prompt to provide comprehensive visual details and auxiliary examples, enhancing LLM robustness in distribution-changing scenarios during inference.
- Extensive experiments on OKVQA, A-OKVQA, VQAv2, VQA-CP, and GQA-OOD demonstrate our method's effectiveness and generalizability, establishing new state-of-the-art performance.

Related Work

Debiasing Methods in VQA

Language bias is one of the most significant issues in VQA, as models are likely to learn easier patterns rather than the target pattern during training. In other words, models tend to remember an unimodal pattern (question-answer correlation) instead of a multi-modal pattern (question-image-answer correlation) during the VQA training (Geirhos et al. 2020). Numerous debiasing methods have been proposed, which can be broadly categorized into five groups: innovative architectures (Anderson et al. 2018); methods with an enhanced Language Model (LM); methods with improved visual information (Selvaraju et al. 2019; Wu and Mooney 2019); ensemble methods (Clark, Yatskar, and Zettlemoyer 2019); data-driven strategies (Chen et al. 2023). Recently, numerous studies have argued that large models still suffer from language bias inherited from their training data. In this work, we aim to alleviate language bias for VQA using LLMs by enhancing visual descriptions from an object attribute perspective.

Knowledge-Based VQA

There is a trend that LLMs are getting increasingly important in KBVQA. In 2022, Yang et al. (Yang et al. 2022) first utilized GPT-3 (Brown et al. 2020) for KBVQA in a few-shot setting, which they referred to as PICa. Building upon PICa, Shao et al. (Shao et al. 2023) further enhanced GPT-3's comprehension of the task by encoding answer heuristics, answer candidates, and answer-aware examples into the prompts. In addition to the few-shot scenarios, Guo et al. (Guo et al. 2023) proposed a zero-shot VQA method called Img2LLM, which utilizes the OPT model (Zhang et al. 2022). Moreover, some efforts have also focused on improving the prompt's design, such as the Reasoning Question Prompts (RQP) proposed by Lan et al. (Lan et al. 2023) for zero-shot VQA, which further promoted the potential of LLMs from a question perspective. Hu et al. (Hu et al. 2023) further advanced this study via a novel captioning model that bridges the gap between image and LLMs. Recently, Zhang et al. (Zhang, Jiang, and Zhao 2024) focused on improving few-shot LLM-based KBVQA via reducing language bias inherited by LLMs. In this work, we aim to enhance the domain-shift capability of LLM-based VQA by incorporating practical memory knowledge in a zero-shot setting.

Methodology

In this section, we present a comprehensive overview of the proposed method. Specifically, we first provide a thorough illustration of the OAD-Promoter architecture, then offer an elaborate explanation of the OEG module and the MKA module, and provide a detailed description of the OAD Prompt.

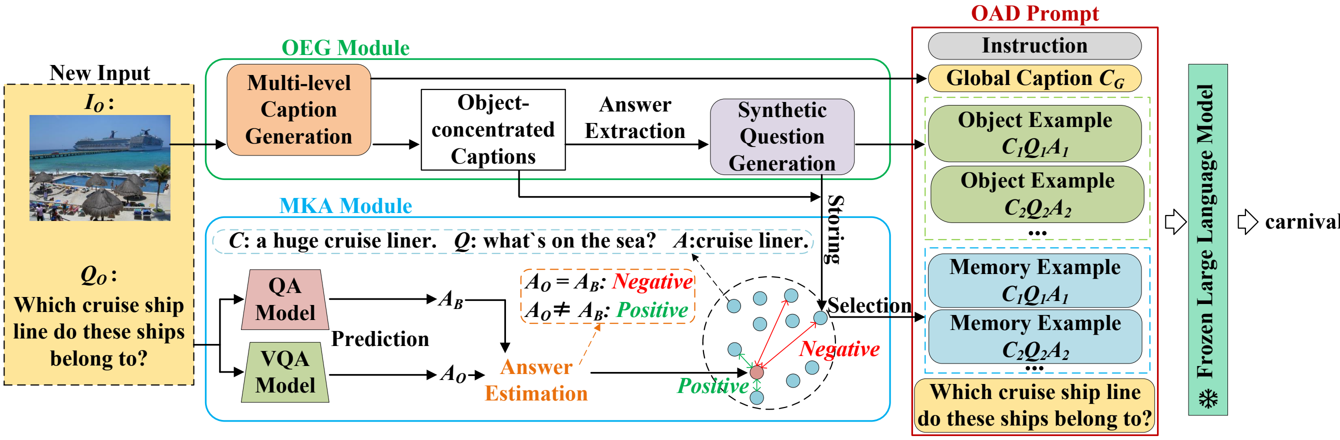

Figure 2: Architecture of OAD-Promoter. It comprises three components: 1) The OEG module (green box) generates a global caption and object-focused samples; 2) The MKA module (blue box) processes novel inputs by leveraging relevant stored examples to assist the LLM; and 3) The OAD Prompt (red box) integrates outputs from the preceding modules and directs the LLM to produce the final answer.

Architecture

We follow the LLM-based VQA pipeline in previous works (Yang et al. 2022; Shao et al. 2023; Guo et al. 2023), the architecture is illustrated in fig. 2, which consists of three components: OEG Module, MKA Module, and OAD Prompt. The OEG Module is responsible for generating a global caption and object-concentrated examples; the MKA module plays a role in assisting LLMs to handle new inputs with the support of relevant stored examples; and the OAD Prompt aims to integrate the output of the former two modules and guide LLMs to predict the output. The proposed method can effectively overcome language bias and improve the LLM's capacity in dealing with OOD samples at the same time.

To be specific, in the OEG module, motivated by visual-enhancing debiasing methods (Selvaraju et al. 2019; Wu and Mooney 2019), we use object-concentrated descriptions to compensate for the absence of fine-grained visual information in the global caption. Thereby, the language bias is mitigated by this visual enhancement approach. Besides, the generated object-concentrated examples are helpful for the LLM in better understanding the detailed visual content of the input image. In the MKA module, to further alleviate language bias and ensure that learned examples remain available in the prompt's architecture, we employ a QA model that outputs a biased answer and a VQA model that conducts an ordinary VQA prediction to pre-assess the language bias of the new input. The computation algorithm depends on the relationship between the outputs (the biased answer and the ordinary answer), and the auxiliary examples are chosen via the similarity computation. In this way, the LLM's exploitation of language bias can be prevented, and its inference on new inputs is strengthened via memory knowledge. In the OAD Prompt, the outputs of the OEG module and the MKA module are integrated into the construction of the prompt. The OAD Prompt's content guides the LLM and predicts the final output. Note that our method's entire procedure does not utilize any external knowledge sources or data that need to be retrieved, making it a pure zero-shot method.

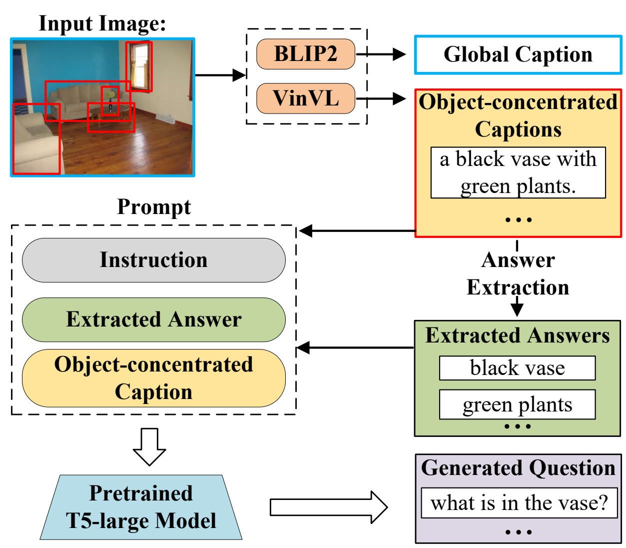

Figure 3: The illustration of the detailed process in the OEG module. Specifically, the BLIP2 and VinVL are used to produce the global caption and object-concentrated captions, respectively. The generated questions are output by a pretrained T5-large model via prompting.

OEG Module

The detailed process in the OEG module is shown in fig. 3. This module comprises two primary generation processes: multi-level caption generation and synthetic question generation. The object-concentrated captions produced by the former process serve as an input to the latter process.

Multi-level Caption Generation. This process is in charge of generating captions according to the input image. For the input image I_O, we utilize a pre-trained BLIP2 (Li et al. 2023) and a VinVL detector (Zhang et al. 2021) to generate a global caption and a couple of regional captions that focus on individual object attributes, respectively.

Synthetic Question Generation. This process is responsible for producing questions according to the generated object-concentrated captions and constitutes complete examples for constructing the prompt. As for an object-concentrated caption, we follow the operation in Img2LLM (Guo et al. 2023), and adopt an effective caption evaluation tool (Lee et al. 2021) to extract answers from the caption. Potential answers are extracted in the form of noun phrases, verb phrases, adjective phrases, numbers, as well as boolean-typed words such as "yes" and "no". After getting potential answers, we utilize a neural question generation method to produce corresponding questions. As illustrated in fig. 3, we form a prompt as: "[$Instruction$ / $Extracted \quad Answers$ / $Object - concentrated \quad Caption$]" to guide a pre-trained T5-large model (Raffel et al. 2020) that finetuned on SQuAD2.0 (Rajpurkar, Jia, and Liang 2018), MultiRC (Khashabi et al. 2018), BookQA (Mihaylov et al. 2018), CommonsenseQA (Talmor et al. 2019) and Social IQa (Sap et al. 2019) to produce questions. Utilizing the above question generation approaches, we obtain a collection of synthetic question-answer (QA) pairs for the subsequent operations.

With the above two processes, an object-concentrated example is formed with the combination of an object-concentrated caption and a synthetic QA pair. Note that each example E_i that contains a caption C, a question Q, and an answer A is utilized as an element both in the stored memory knowledge of the MKA module and in the prompt construction of the OAD Prompt.

MKA Module

The MKA module is used to assist LLM-based VQA in tackling new inputs by leveraging the memory knowledge from stored object-concentrated examples and supporting the LLM in making more reliable predictions in distribution-changing circumstances. This module comprises two primary processes: answer estimation and similarity computation. The former process aims to estimate the language bias of the new input, while the latter process is responsible for selecting the appropriate stored examples for the subsequent construction of the OAD Prompt.

Answer Estimation. We use a general VQA model (Anderson et al. 2018) and an off-shift QA model in LMH (Clark, Yatskar, and Zettlemoyer 2019) in the language bias assessment of the new input. As illustrated in fig. 2, when the input image and question are sent to the MKA module, the VQA model and the QA model both make predictions depending on the given image and question. Then, the answers output by the two models are estimated before the similarity computation starts. Concretely, since the QA model's inference involves no visual information, the result from it is a biased answer, while the result from the VQA model is ordinary. Let A_B and A_O be the output from the QA model and the VQA model, respectively. The selection mode M is defined by the Eq. (1):

$$M = \begin{cases} Positive, & \text{if } A_O \neq A_B; \\ Negative, & \text{if } A_O = A_B. \end{cases} \tag{1}$$

Note that the situation of A_B = A_O means the ordinary VQA model utilizes the language bias during inference, and this implies the LLM can exploit this language bias as well because the scale of the LLM's training is much larger than the ordinary model's training. Therefore, there is a higher probability that the LLM takes advantage of the language bias than a general VQA model does. Hence, in this situation, we adopt a "Negative" selection mode that conflicts with the biased result (A_B or A_O) in the subsequent similarity computation. On the contrary, if $A_B \neq A_O$, that means there is no sign of the language bias that could be used in the prediction, we choose a "Positive" selection mode that supports the ordinary result A_O for enhancing the LLM's inference.

Similarity Computation. Inspired by Prophet (Shao et al. 2023), as for the new input and every stored objective-concentrated example E_i, we first extract the fused visual-language feature from the image and the question. Then, the feature f of the new input and the feature f_i of the E_i are the elements in the cosine similarity computation. To be specific, the f is extracted from the aforementioned VQA model in the last process. The visual feature of the f_i comes from the object attribute image of the VinVL, and the language feature comes from the generated question of the pre-trained T5-large model; these two modal features are encoded by the VQA model as well, and f_i is the output. After obtaining the fused features, the cosine similarity is computed by Eq. (2):

$$E_S = \begin{cases} argTopN \frac{f^T f_j}{\|f\|2\|f_j\|2}, & \text{if } M = Positive; \\ argBottomN \frac{f^T f_j}{\|f\|2\|f_j\|2}, & \text{if } M = Negative, \end{cases} \tag{2}$$

where $i \in \{1, 2, \ldots, K\}$.

The E_S is an index set of the N selected examples in the memory latent space; K is the number of stored examples; M is the selection mode. Note that if M = Positive, we choose the most similar examples in the memory latent space; otherwise, the least similar examples would be selected.

OAD Prompt

The OAD Prompt is responsible for integrating outputs from the former two modules and leading the inference of the LLM. Previous works (Jin et al. 2022; Lan et al. 2023) have proved that the design of prompts is critical in the LLM's inference. Although existing methods have achieved remarkable results in the few-shot or zero-shot settings, they do not account for both global image description and object attribute description. The function of the OAD Prompt includes:

- Providing more comprehensive and more meticulous image information to LLMs, thereby mitigating language bias by enhanced visual information.
- Giving similar memory examples for LLMs when it handles the new input. Thus strengthening the reliability of its prediction with the memory knowledge assistance.
- As the inference proceeds, the number of memory examples in the latent space is growing, and the auxiliary

memory knowledge is getting more massive and multifarious. Therefore, the domain-shift capability of LLMs is facilitated constantly.

Different from the prompt style in previous works (Yang et al. 2022; Shao et al. 2023; Guo et al. 2023; Lan et al. 2023), the proposed OAD Prompt considers both the global caption and object attribute description of the input image, which is typically overlooked by existing LLM-based VQA methods. The architecture of the OAD Prompt is illustrated in fig. 2, which comprises four parts: the instruction I; the global caption C_G; the examples E; and the inputted question Q_O. Note that E contains two kinds of examples: object-concentrated examples E_O that are derived from the input image and memory examples E_S that are selected from the memory latent space, and each example can be formulated as (C, Q, A), which represents a triple of a caption, a question, and an answer.

Experiments

Datasets and Metrics

We evaluate the proposed method on three benchmark datasets: VQAv2 (Goyal et al. 2017), OKVQA (Marino et al. 2019), and A-OKVQA (Schwenk et al. 2022), where the image content is not sufficient to answer the questions. Thus, models have to perform reasoning based on the perception of even common sense to answer. Specifically, OKVQA is a commonly used dataset for KBVQA, which contains 9K and 5K image-question pairs for training and testing, respectively. A-OKVQA, an augmented successor of OKVQA, and its questions are more challenging, containing 17K, 1K, and 7K image-question pairs for training, validation, and testing, respectively. Moreover, we utilize VQA-CP (Agrawal et al. 2018) and GQA-OOD (Kervadec et al. 2021) to test the generalization capabilities of the models on unseen data. VQA-CP evaluates the performance of the models on data beyond what they were trained on. Meanwhile, GQA-OOD consists of separate head and tail sets, which are designed for the models to conduct reasoning in unfamiliar scenarios. We use soft accuracy (Goyal et al. 2017) as the metric in our experiments.

Implementation Details

As for the general VQA model, we utilize a lightweight yet effective UpDn (Anderson et al. 2018) for encoding fused features and answer estimation. Similar to previous works (Shao et al. 2023; Marino et al. 2021), we further enhance the model's capabilities by implementing the transfer learning paradigm. Initially, the model undergoes pre-training on the VQAv2 dataset and the Visual Genome dataset (Krishna et al. 2017). To avoid data pollution, we eliminate samples from the pre-training dataset if their images are utilized in the testing split of OKVQA. Subsequently, the pre-trained model is further finetuned on the training split of OKVQA to acquire our final VQA model. Regarding the prompt's structure, since the memory module is empty at the beginning, the initial design of the OAD prompt is formed as follows: "[I / C_G / E_O / Q_O]". And this design turns into "[I / C_G / E_O / E_S / Q_O]" when the next sample inputted. As for LLMs, we adopt GPT-3 (Brown et al. 2020) and OPT (Zhang et al. 2022) as frozen LLMs in our main experiments for a fair comparison with previous methods (i.e., PICa, Prophet, Img2LLM, PromptCap, GRACE). Moreover, we also use other LLMs in the experiment to validate the generalization ability of our method.

Quantitative Results

We conduct comparative experiments with various KBVQA methods and methods that need large-scale vision-language pretraining. All methods used in experiments follow their official instructions. The main quantitative results are summarized in table 1. **Compared with the methods with large-scale multi-modal pretraining,** OAD-Promoter shows remarkable performance that is superior to all of them. **Compared with the methods with frozen LLMs,** OAD-Promoter yields competitive results both in few-shot and zero-shot settings. Especially, a new state-of-the-art result was attained on the VQAv2 dataset under a zero-shot scenario.

We also compare OAD-Promoter with the latest method (GRACE (Zhang, Jiang, and Zhao 2024)) using different LLMs (i.e., LLaVA-1.5 (Liu et al. 2024), LLaMA2 (Touvron et al. 2023), and GPT-4 (Brown et al. 2020)) in few-shot scenario and report results in table 2. Although our results are inferior to those of LLaVA-1.5 and LLaMA2, our method achieves the best accuracy on GQA-OOD with GPT-4, which suggests that OAD-Promoter can demonstrate a better domain-shift capacity with the more advanced frozen LLM.

For further validation of the zero-shot ability and generalization capability of our methods, we use different sizes of diverse LLMs (i.e., GPT-3 (Brown et al. 2020), OPT (Zhang et al. 2022), BLOOM (Scao et al. 2022), GPT-Neo (Black et al. 2021), GPT-J (Wang and Komatsuzaki 2021)) in zero-shot evaluation on OKVQA TEST SET with other zero-shot KBVQA methods, the results are presented in table 3. These results highlight that the proposed method can integrate various existing LLMs, which proves the generalization competence of our method.

Moreover, as illustrated in table 4, to validate the effectiveness of the proposed method compared with the debiasing integration, we conduct a few-shot comparison experiment to compare three methods with the debiasing methods (Clark, Yatskar, and Zettlemoyer 2019; Chen et al. 2023) that can be integrated into the LLM-based pipeline. From the accuracy of OKVQA, we can see the result drops with the integration of debiasing methods. This phenomenon demonstrates that LLM-based methods tend to exploit language bias when addressing knowledge-intensive questions, and it suggests that there is indeed a lack of reliability in the answer output by LLMs as well.

Qualitative Analysis

The qualitative analysis is shown in fig. 4. The representative zero-shot KBVQA method (Img2LLM (Guo et al. 2023)) is compared with the OAD-Prompter, and four cases are shown here. Note that cases (a-b) are from close domains concerned with sport and athletics; however, case (c) comes from a different domain that is about electronic devices or

Method	Few/Zero-shot	VQAv2		A-OKVQA		OKVQA
		val	test	val	test	test
Methods with Large-scale Multi-modal Pretraining						
VL-T5 (Cho et al. 2021)	Zero-shot	13.50	14.13	-	-	5.73
Frozen (Tsimpoukelli et al. 2021)	Zero-shot	29.43	29.55	-	-	5.90
Flamingo-3B (Alayrac et al. 2022)	Zero-shot	48.24	49.18	-	-	41.17
Flamingo-9B (Alayrac et al. 2022)	Zero-shot	50.77	51.80	-	-	44.64
Flamingo-80B (Alayrac et al. 2022)	Zero-shot	56.08	56.21	-	-	50.57
FewVLM-base (Jin et al. 2022)	Zero-shot	43.28	43.39	-	-	11.52
FewVLM-large (Jin et al. 2022)	Zero-shot	47.68	47.73	-	-	16.50
VLKD-ViT-B/16 (Dai et al. 2022)	Zero-shot	38.60	39.69	-	-	10.50
VLKD-ViT-L/14 (Dai et al. 2022)	Zero-shot	42.55	44.48	-	-	13.24
Methods with Frozen Large Language Models						
PICa (Yang et al. 2022) *w/* GPT-3	Few-shot	56.10	56.12	47.08	47.64	48.01
REVIVE (Lin et al. 2022) *w/* GPT-3	Few-shot	57.11	57.69	57.33	57.05	58.03
KAT (Gui et al. 2022) *w/* GPT-3	Few-shot	56.30	56.48	55.38	55.12	54.41
Prophet (Shao et al. 2023)*w/* GPT-3	Few-shot	58.37	59.22	**59.27**	**57.30**	**61.08**
PromptCap (Hu et al. 2023) *w/* GPT-3	Few-shot	58.18	58.77	58.86	57.28	60.44
GRACE (Zhang, Jiang, and Zhao 2024) *w/* GPT-3	Few-shot	58.07	58.45	58.61	57.15	60.29
OAD-Promoter *w/* GPT-3	Few-shot	57.96	58.42	58.50	56.99	60.04
PICa (Yang et al. 2022) *w/* GPT-3	Zero-shot	28.67	29.30	23.74	26.88	17.63
PICa (Yang et al. 2022) *w/* GPT-3 + RQP (Lan et al. 2023)	Zero-shot	28.70	29.34	28.92	27.73	20.27
Img2LLM (Guo et al. 2023) *w/* GPT-3	Zero-shot	58.69	59.22	38.88	43.39	42.80
Img2LLM (Guo et al. 2023) *w/* GPT-3 + RQP (Lan et al. 2023)	Zero-shot	58.96	59.35	42.88	43.61	45.57
Img2LLM (Guo et al. 2023) *w/* OPT	Zero-shot	60.58	61.83	42.90	40.69	45.58
Img2LLM (Guo et al. 2023) *w/* OPT + RQP (Lan et al. 2023)	Zero-shot	60.60	61.85	42.83	40.64	45.52
OAD-Promoter *w/* OPT	Zero-shot	60.62	61.93	43.03	40.68	45.58
OAD-Promoter *w/* GPT-3	Zero-shot	**60.64**	**61.98**	43.09	41.71	45.61

Table 1: Performance comparison on VQAv2, A-OKVQA, and OKVQA TEST SET. The bold fonts indicate the best results in the entire table, and the underlined fonts denote the best results among methods with frozen LLMs in the zero-shot setting. Note that we use GPT-3 in this experiment for the sake of a fair comparison.

LLM	GRACE		OAD-Promoter	
	VQA-CP	GQA-OOD	VQA-CP	GQA-OOD
LLaVA-1.5	**54.08**	**48.96**	53.94	48.93
LLaMA2	**57.32**	**50.20**	55.89	49.67
GPT-4	**57.61**	50.19	55.93	**50.21**

Table 2: Few-shot performance comparison on VQA-CP and GQA-OOD with different LLMs. The best results are highlighted in bold and underlined.

Method	LLM Size	OKVQA
Frozen	7B	5.90
PICa *w/* GPT-3	175B	17.63
Img2LLM *w/* OPT	6.7B	38.20
Img2LLM *w/* OPT	30B	41.82
Img2LLM *w/* OPT	175B	45.58
OAD-Promoter *w/* BLOOM	7.1B	33.77
OAD-Promoter *w/* OPT	6.7B	36.18
OAD-Promoter *w/* OPT	30B	40.46
OAD-Promoter *w/* OPT	175B	45.58
OAD-Promoter *w/* GPT-Neo	2.7B	33.41
OAD-Promoter *w/* GPT-J	6B	38.89
OAD-Promoter *w/* GPT-3	175B	45.61

Table 3: Zero-shot evaluation of the OAD-Promoter based on different sizes of diverse LLMs.

technology brands, and case (d) derives from a highly distinct domain in language. These cases are sent to the two methods in alphabetical order. Our method achieves 100% correctness in four cases, although these cases involve three different domains. The Img2LLM attains 75% accuracy, but it fails in case (d). In addition, we reverse the input order of these four cases and observe the results of the two methods. As a result, our method still maintains 100% correctness, and Img2LLM also achieves all right predictions. This interesting situation uncovers that the input order also has an influence on the LLM's inference when it deals with samples from multiple different domains The experiment confirms that our method exhibits better adaptation capability in a distribution-changing environment.

Ablation Study

To assess the contributions of the proposed OEG and MKA modules, we conduct ablation studies by evaluating our method with and without each module. Note that we use

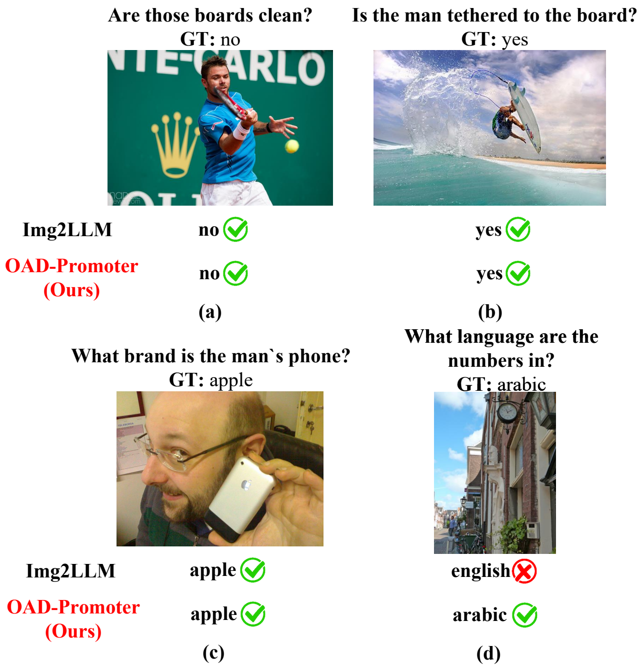

Figure 4: Qualitative analysis of the proposed method. Four cases from distinct domains are displayed.

Method	OKVQA	VQA-CP
PICa	48.01	41.90
PICa + LMH	47.91 ↓	43.13
PICa + LMH + CSS	46.06 ↓	46.87
Prophet	**61.11**	53.41
Prophet + LMH	59.96 ↓	54.33
Prophet + LMH + CSS	57.28 ↓	55.31
GRACE	60.32	57.35
GRACE + LMH	59.92 ↓	58.72
GRACE + LMH + CSS	60.15 ↓	**61.37**
OAD-Promoter	60.04	56.47

Table 4: Few-shot evaluation compared with KBVQA methods with debiasing strategies LHM (Clark, Yatskar, and Zettlemoyer 2019) and CSS (Chen et al. 2023).

the example integration method in PICa and Img2LLM to substitute the OEG module in the few-shot and zero-shot settings, respectively. The results in table 5 demonstrate that both the OEG and MKA modules jointly contribute to performance improvements in our architecture. Moreover, as illustrated in table 6, to probe the influence of example number K on the accuracy as the inference proceeds, we manually add examples in a fixed number in the MKA module at the beginning. Note that, unlike the few-shot setting, we provide numerous external examples for the MKA module at the very beginning of the inference. The results reveal that the more examples are provided, the better the performance.

To explore the effect of prompt design on the LLM's inference, we conduct experiments with two designs of the

Method	OEG	MKA	OKVQA
Few-shot setting			
OAD-Promoter	○	○	47.33
OAD-Promoter	●	○	54.68
OAD-Promoter	○	●	48.95
OAD-Promoter	●	●	60.04
Zero-shot setting			
OAD-Promoter	○	○	42.50
OAD-Promoter	●	○	44.26
OAD-Promoter	○	●	43.64
OAD-Promoter	●	●	45.61

Table 5: Ablation study on the proposed OEG and MKA modules on OKVQA TEST SET.

Method	OEG	MKA	K	OKVQA
OAD-Promoter	○	●	0	43.64
OAD-Promoter	○	●	60	43.65
OAD-Promoter	○	●	200	43.92
OAD-Promoter	○	●	400	44.15

Table 6: Ablation study on the effect of example number in the MKA module. Note that we introduce K examples into the MKA module manually at the beginning.

Prompt Design	OKVQA
CCC-QAQAQA	44.82
CQA-CQA-CQA	45.61

Table 7: Ablation study on the OAD Prompt with two prompt designs on OKVQA TEST SET under zero-shot setting.

OAD Prompt and report their results in table 7. The C, Q, and A stand for context, question, and answer in one example, respectively. Assuming the number of examples is 3, the first design is $CCC - QAQAQA$, the second is $CQA - CQA - CQA$. The results illustrated that a complete entirety, which contains a context, a question, and an answer, is more helpful for the OAD Prompt.

Conclusion

In this work, we propose OAD-Promoter, a novel zero-shot KBVQA method designed to mitigate language bias and strengthen domain-shift robustness in LLMs by providing multi-granularity captions and auxiliary stored examples as enhanced information to LLMs. By integrating three components: OEG module, MKA module, and OAD Prompt, OAD-Promoter achieves competitive performance compared to few-shot LLM-based methods and establishes a new state-of-the-art among zero-shot LLM-based approaches. Comprehensive quantitative and qualitative experiments on OKVQA, A-OKVQA, VQAv2, VQA-CP, and GQA-OOD datasets validate its effectiveness. We believe OAD-Promoter can contribute to advancing LLM-based VQA research in zero-shot and real-world applications.

Acknowledgments

This work was in part supported by the Science and Technology Development Fund of Macau, Macau SAR (Grants No. 0035/2023/ITP1 and 0021/2023/RIA1); the National Natural Science Foundation of China (Grant No. 62376196); and the Tianjin Natural Science Foundation (Grant No. 24JCJQJC00190).

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4971–4980.

Alayrac, J.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 23716–23736.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2425–2433.

Black, S.; Gao, L.; Wang, P.; Leahy, C.; and Biderman, S. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. https://doi.org/10.5281/zenodo.5297715.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. In *Proceedings of the International Conference on Neural Information Processing Systems*, 1877–1901.

Chen, L.; Zheng, Y.; Niu, Y.; Zhang, H.; and Xiao, J. 2023. Counterfactual Samples Synthesizing and Training for Robust Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13218–13234.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying Vision-and-Language tasks via text generation. In *Proceedings of the International Conference on Machine Learning*, 1931–1942.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 4067–4080.

Dai, W.; Hou, L.; Shang, L.; Jiang, X.; Liu, Q.; and Fung, P. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. In *Findings of the Annual Meeting of the Association for Computational Linguistics*, 2383–2395.

Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Goyal, Y.; Khot, T.; Agrawal, A.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6904–6913.

Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; and Gao, J. 2022. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 956–968.

Guo, J.; Li, J.; Li, D.; Tiong, A. M. H.; Li, B.; Tao, D.; and Hoi, S. 2023. 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*, 10867–10877.

Hu, Y.; Hua, H.; Yang, Z.; Shi, W.; Smith, N. A.; and Luo, J. 2023. PromptCap: Prompt-Guided Task-Aware Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2963–2975.

Jin, W.; Cheng, Y.; Shen, Y.; Chen, W.; and Ren, X. 2022. A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2763–2775.

Kervadec, C.; Antipov, G.; Baccouche, M.; and Wolf, C. 2021. Roses are red, violets are blue... but should VQA expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2776–2785.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 252–262.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1): 32–73.

Lan, Y.; Li, X.; Liu, X.; Li, Y.; Qin, W.; and Qian, W. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the ACM International Conference on Multimedia*, 4389–4400.

Lee, H.; Scialom, T.; Yoon, S.; Dernoncourt, F.; and Jung, K. 2021. QACE: Asking Questions to Evaluate an Image Caption. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, 4631–4638.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the International Conference on Machine Learning*, 19730–19742.

Li, P.; Si, Q.; Fu, P.; Lin, Z.; and Wang, Y. 2024. Object Attribute Matters in Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18545–18553.

Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; and Yuan, L. 2022. REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering. In *Proceedings of the International Conference on Neural Information Processing Systems*, 10560–10571.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; and Rohrbach, M. 2021. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14111–14121.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3195–3204.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2381–2391.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 784–789.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019. SocialIQa: Commonsense reasoning about social interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 4463–4473.

Scao, T. L.; Wang, T.; Hesslow, D.; Bekman, S.; Bari, M. S.; Biderman, S.; Elsahar, H.; Muennighoff, N.; Phang, J.; Press, O.; et al. 2022. What Language Model to Train if You Have One Million GPU Hours? In *Findings of the Conference on Empirical Methods in Natural Language Processing*, 765–782.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision*, 146–162.

Selvaraju, R. R.; Lee, S.; Shen, Y.; Jin, H.; Ghosh, S.; Heck, L. P.; Batra, D.; and Parikh, D. 2019. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2591–2600.

Shao, Z.; Yu, Z.; Wang, M.; and Yu, J. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14974–14983.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 4149–4158.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Tsimpoukelli, M.; Menick, J.; Cabi, S.; Eslami, S. M. A.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. In *Proceedings of the International Conference on Neural Information Processing Systems*, 200–212.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax.

Wu, J.; and Mooney, R. J. 2019. Self-Critical Reasoning for Robust Visual Question Answering. In *Proceedings of the International Conference on Neural Information Processing Systems*, 8604–8614.

Yang, L.; Xiao, Z.; Huang, W.; and Zhong, X. 2025. StoryLLaVA: Enhancing Visual Storytelling with Multi-Modal Large Language Models. In *Proceedings of the International Conference on Computational Linguistics*, 3936–3951.

Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; and Wang, L. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3081–3089.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M. T.; Li, X.; Lin, X. V.; et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*. https://doi.org/10.48550/arXiv.2205.01068.

Zhang, X.; Zhang, F.; and Xu, C. 2023. VQACL: A Novel Visual Question Answering Continual Learning Setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19102–19112.

Zhang, Y.; Jiang, M.; and Zhao, Q. 2024. GRACE: Graph-Based Contextual Debiasing for Fair Visual Question Answering. In *Proceedings of the European Conference on Computer Vision*, 176–194.

Zhong, X.; Li, Z.; Chen, S.; Jiang, K.; Chen, C.; and Ye, M. 2023. Refined Semantic Enhancement towards Frequency Diffusion for Video Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3724–3732.