Enhanced Reasoning via Multimodal LLMs and Collaborative Inference

Zhiquan Wen, Mingkui Tan, *Member, IEEE*, Yaowei Wang, *Member, IEEE*, Qingyao Wu, and Qi Wu

***Abstract*—Visual Question Answering (VQA) is a prevalent task that can facilitate the perception of the real world by the visually impaired. However, many VQA models tend to rely on superficial correlations in datasets for predictions rather than genuine reasoning, limiting their real-world applicability. While existing methods address this issue by incorporating debiasing strategies during training, they typically assume prior knowledge of out-of-distribution (OOD) test sets and then tailor debiasing strategies and select optimal models on the basis of the OOD samples. This reliance on OOD test data, however, is unrealistic in practical applications. To address this, some works introduce test-time adaptation techniques to mitigate dataset shifts during model deployment. Despite their potential, these methods risk catastrophic forgetting as they update models at test time without access to the ground-truth answers or the source data. An emerging solution involves leveraging the extensive knowledge embedded in Large Language Models (LLMs) to support reasoning tasks, yet their language-only input restricts flexibility in multimodal tasks. To bridge this gap, we propose leveraging the zero-shot capability of Multimodal Large Language Models (MLLMs). To optimise computational efficiency, we introduce a novel VQA Collaborative Inference framework (VQA-CI) that integrates MLLMs (e.g., BLIP-2 Flan T5) with VQA specialists (e.g., UpDn). This framework initially processes samples through VQA specialists and subsequently determines the necessity for re-evaluation with MLLMs based on predefined bias and reliability indicators. Experiments on the GQA-OOD and VQA-CP v2 datasets show that our VQA-CI achieves significant performance gains, with accuracy improvements of around 6% over state-of-the-art methods, underscoring the effectiveness of our VQA-CI.**

***Index Terms*—Collaborative inference, multimodal large language model, zero-shot visual question answering.**

I. Introduction

VISUAL Question Answering (VQA) [1], [2], [3], [4], [5] has emerged as a cutting-edge interdisciplinary field at the confluence of Computer Vision (CV) [6], [7], [8] and Natural Language Processing (NLP) [9], [10], [11], aiming to endow computational systems with the ability to comprehend visual content in images and accurately respond to questions posed in natural language. An exceptional VQA system possesses advanced reasoning capabilities to interpret questions accurately, thereby enabling visually impaired individuals to bridge the gap between their perception of the environment and the visual world. While certain works [1], [2] have demonstrated remarkable performance on benchmark datasets (e.g., VQA v2 dataset [12]), recent studies [13], [14] have revealed that numerous VQA models tend to leverage shortcuts, relying on direct associations between unimodal information and answers for predictions, rather than real reasoning capabilities. As shown in Fig. 1, the UpDn model, trained on a biased training set (e.g., where "2" dominates the answer distribution), tends to rely on the learned biases to answer questions, instead of the reasoning capabilities. Specifically, it often predicts "2" while ignoring the image information, resulting in poor generalisation, as evidenced by the predicted answer distributions of the UpDn model, which closely resemble the training set distribution but differ significantly from the test set distribution. This reliance on superficial connections (i.e., bias) significantly constrains the practical applicability of VQA models in real-world scenarios.

To overcome bias issues, most contemporary methods [16], [17], [18] adopt debiasing strategies during the training phase and select the best model on the basis of out-of-distribution (OOD) test sets (e.g., VQA-CP v2 dataset [15]). For example, Niu et al. [18] leverage cause-effect analysis to comprehensively examine language bias and propose a counterfactual inference framework to mitigate bias issues effectively. Moreover, Wen et al. [16] contend that biases permeate both vision and language modalities and seek to overcome bias issues from feature and sample perspectives. However, the unpredictability of OOD distributions in real-world scenarios, which remain unknown until the model's evaluation phase, presents a substantial obstacle. This reality makes the application of training-time debiasing techniques to real-world settings considerably challenging. To circumvent this limitation, TDS [19] introduces a novel test-time debiasing approach that adopts self-supervised and unsupervised objectives (e.g., entropy) to update models. Despite this innovation, the approach risks inducing catastrophic forgetting. Benefiting from pre-training on extensive and diverse corpora, Large Language Models (LLMs) [20], [21], [22], [23] inherently cover a wide range of data distributions, potentially mitigating bias issues. The formidable zero-shot

Received 31 July 2024; revised 11 December 2024 and 18 January 2025; accepted 24 January 2025. Date of publication 21 July 2025; date of current version 21 October 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U24A20327 and Grant 62072190, and in part by the Major Key Project of Peng Cheng Laboratory (PCL) under Grant PCL2023A08. The associate editor coordinating the review of this article and approving it for publication was Prof. Ngai-Man Cheung. *(Corresponding author: Mingkui Tan.)*

Zhiquan Wen is with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China, and also with PengCheng Laboratory, Shenzhen 518055, China (e-mail: sewenzhiquan@mail.scut.edu.cn).

Mingkui Tan and Qingyao Wu are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: mingkuitan@scut.edu.cn; qyw@scut.edu.cn).

Yaowei Wang is with PengCheng Laboratory, Shenzhen 518055, China (e-mail: wangyw@pcl.ac.cn).

Qi Wu is with the School of Computer Science, University of Adelaide, Adelaide 5005, Australia (e-mail: qi.wu01@adelaide.edu.au).

Digital Object Identifier 10.1109/TMM.2025.3590940

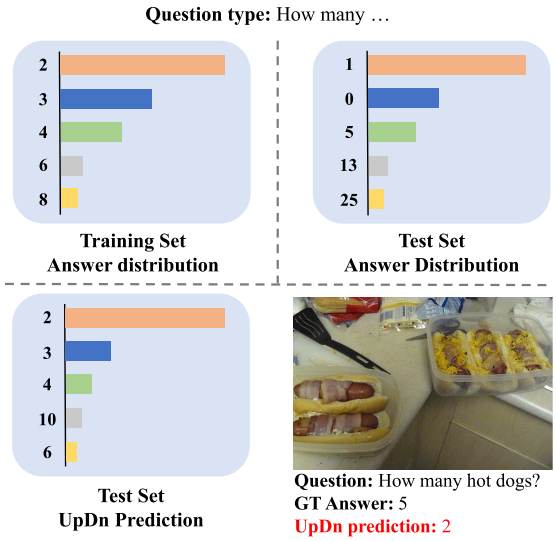

Fig. 1. Answer distribution for "How many... " questions on the VQA-CP v2 [15] dataset. In the training set, "2" dominates the answer distribution, whereas in the test set, "1" is the most frequent answer, with "2" appearing rarely. After training on the biased training set, the UpDn [2] model tends to rely on these learned "biases" to answer questions rather than employing reasoning abilities.

reasoning capabilities of LLMs may present a promising avenue for addressing bias concerns. For example, PICa [24] addressed the VQA task by transforming images into captions and tags, leveraging in-context learning with GPT-3 [22] to generate answers. However, the inherent limitation of LLMs to process exclusively linguistic input poses challenges for multimodal reasoning tasks.

To address the aforementioned limitations, Multimodal Large Language Models (MLLM) [25], [26], [27] have emerged. MLLMs incorporate mechanisms such as linear layers [28] or the Q-Former [25] to convert information from other modalities into tokens comprehensible to LLMs, thereby enhancing the flexibility of LLMs in handling multimodal reasoning tasks. However, the vast parameter space of MLLMs and the scarcity of training data for downstream tasks render fine-tuning MLLMs for specific applications impractical. Consequently, we seek to leverage MLLMs' zero-shot capabilities to circumvent bias issues and facilitate more reliable reasoning. Nonetheless, the process of transforming information from other modalities into tokens may lead to the omission of vital information, potentially constraining the effectiveness of MLLMs' zero-shot reasoning. To mitigate this challenge, Awal et al. [29] explored the integration of general image captions as supplemental information to help the reasoning process of MLLMs. Their findings suggest that while general image captions can provide valuable context, they may also introduce question-unrelated information, adversely affecting the zero-shot performance of MLLMs.

To generate question-related image captions, our initial step involves identifying regions within images that are pertinent to the posed questions. We seek to leverage the pre-trained baseline models (e.g., UpDn [2]) to pinpoint these targeted regions. Our observations in Fig. 2 reveal that although VQA models exhibit decreased performance on OOD datasets in terms of top-1 accuracy, they display superior performance in the top-3 and even top-5 accuracy metrics. This phenomenon suggests that VQA models have the potential to answer questions correctly and thus possess the ability to extract discriminative information from datasets. Therefore, employing the attention weights from pre-trained baseline models to identify question-related image regions appears to be reasonable. We subsequently employ a sophisticated image captioning model, such as PromptCap [31], to generate captions for these identified regions. Moreover, to streamline the complexity inherent in VQA tasks, we propose transforming the VQA task into an "answer re-ranking" task. This entails extracting the top-k answers predicted by the pre-trained baseline model and subjecting them to MLLMs for reasoning. By adopting these techniques, we circumvent bias issues and achieve superior reasoning performance in VQA tasks.

Moreover, employing MLLMs for reasoning across all samples is time-consuming and resource-intensive. To mitigate this, we introduce a novel VQA reasoning framework that facilitates collaborative inference by integrating MLLMs with VQA specialists (e.g., UpDn [2]). Specifically, we initially process the samples through VQA specialists. For samples that exceed these models' capabilities, we escalate the processing to MLLMs. To accurately identify samples that surpass the capabilities of VQA specialists, we have developed two indicators: biased and unreliable sample indicators. Samples flagged as either unreliable or biased are subsequently processed by the MLLMs. This reasoning framework effectively harnesses the strengths of VQA specialists and MLLMs, thereby not only mitigating bias within VQA tasks but also enhancing the reliability of the reasoning process, to some extent. Furthermore, this framework can be seamlessly integrated into a cloud-edge computing architecture, where the VQA specialists are deployed at the edge, and the more computationally intensive MLLMs are utilised in the cloud, improving the availability in real-world applications.

Our contributions can be summarised as follows:

- We propose a VQA reasoning framework that integrates MLLMs and VQA specialists for collaborative inference. This approach improves model performance and optimises computational efficiency to some extent.
- We devise biased and unreliable sample indicators to identify samples that exceed the capabilities of VQA specialists. These samples are re-processed with MLLMs, enhancing the reliability of the inference process.
- We propose mining discriminative information from VQA specialists, i.e., question-related image regions and candidate answers, to assist the reasoning of MLLMs.

II. Related Works

A. Overcoming Biases in VQA

Bias issues in VQA [13], [14], [32] often indicate that models may rely on superficial correlations within datasets for predictions, rather than actual reasoning abilities. In other words, some VQA methods [4], [33], [34] achieve promising performance on in-domain datasets (e.g., VQA v2 dataset [12]) but may experience severe performance degradation on out-of-distribution datasets (e.g., VQA-CP dataset [15]). To address this, current methods mainly introduce debiased techniques during training time [16], [17], [18], [32], [35],

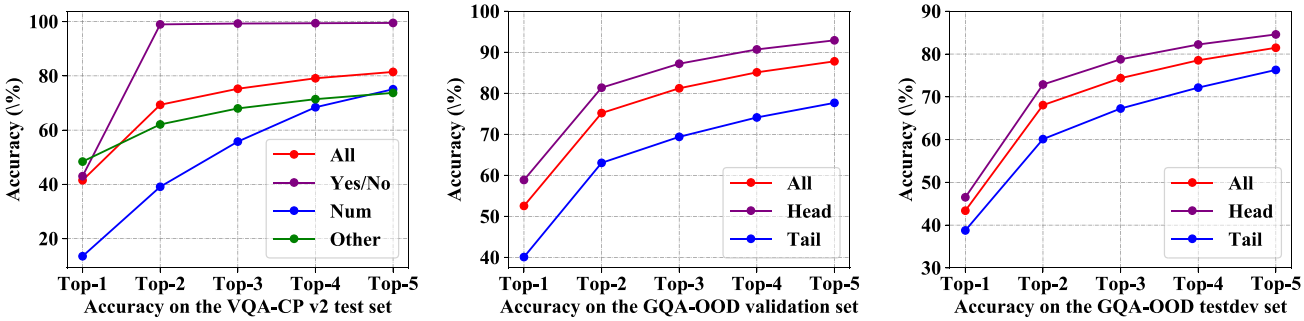

Fig. 2. Performance evaluation of UpDn [2] on the VQA-CP v2 test set [15] and GQA-OOD validation and testdev sets [30] in terms of the top-k accuracy.

[36], [37], [38], [39]. For example, some methods [35], [40] introduce human annotations (such as human attention maps or textual human explanations) to guide VQA models towards recognising essential image regions. Moreover, certain methods establish unimodal-answer branches to identify biases, followed by employing an adversarial learning mechanism [14], dynamically re-weighting samples [32], and adopting cause-effect analysis [18], [41] to mitigate bias issues. Unlike these methods, numerous methods begin with the data's inherent characteristics, constructing additional samples by masking critical objects and words [42], [43], re-composing given images and questions [44], [45], and randomly sampling images or questions to generate negative samples [17], [46], aiming to reduce bias. However, existing training-time methods typically select optimal models using out-of-distributions (OOD) test sets, which are unavailable until real-world evaluation. Recently, advanced Test-Time Adaptation (TTA) techniques in computer vision have provided a new perspective to alleviate this dilemma. Specifically, Gao et al. [47] introduced a test-time prompt tuning method that generates input-adaptive prompts, enabling vision-language models to better recognise novel classes. Beyond updating prompts to the textual branch, Zhang et al. [48] proposed a prototype-based test-time adaptation technique for vision-language models (e.g., CLIP [49]), which dynamically refines prototype features by accumulating task-specific knowledge across modalities. Moreover, Luo et al. [50] devised a two-stage few-shot test time adaptation framework that first fine-tunes a pre-trained source model using a few-shot support set with feature diversity augmentation, followed by model adaptation guided by a prototype memory bank to generate reliable pseudo-labels. Inspired by TTA techniques, Wen et al. [19] introduced a test-time adaptation technique [51], [52], [53] into the VQA domain, devising self-supervised or unsupervised objectives to adapt VQA models at test time to counteract bias issues. Nevertheless, updating models at the test time may lead to catastrophic forgetting. In this paper, we explore overcoming bias issues by leveraging the zero-shot capabilities of Multimodal Large Language Models (MLLMs), offering a novel perspective on the debiasing of VQA.

B. Zero-/Few-Shot Visual Question Answering

In the pursuit of achieving zero- or few-shot VQA, mainstream methods [54], [55], [56], [57], [58] incorporate the superior reasoning capabilities of Large Language Models (LLMs) [22], [23]. Specifically, PICa [24] pioneers this direction by transforming images into captions and tags, using in-context learning with GPT-3 [22] to facilitate comprehension. Building on this, PromptCap [31] refines the process through a question-guided captioning model that is fine-tuned to produce captions relevant to the posed questions. Prophet [59] identified the limitations of vague captions and introduced the concept of providing answer candidates along with confidence scores to guide the prediction process more effectively. Img2LLM [57] enhances the learning paradigm by generating additional question-answer pairs for each instance, simulating a few-shot learning environment. Qivx [58] simplifies the reasoning process for MLLMs by generating sub-questions that reduce the complexity of the primary question. Moreover, Awal et al. [29] delved into the zero- and few-shot capabilities of MLLMs, underscoring the potential of these models in VQA. Inspired by these methods, we employ MLLMs as the foundation for zero-shot reasoning within our framework. Our method enhances this foundation by generating captions relevant to the questions. Moreover, we regard the VQA task as an "answer re-ranking" task. Here, we extract the top-k answers from VQA specialists (e.g., UpDn) on the basis of their probability and then refine the selection of answers by exploiting the sophisticated reasoning capabilities inherent in MLLMs.

C. Multimodal Large Language Models

Large Language Models (LLMs) [20], [21], [22], [23] contain rich knowledge and demonstrate profound reasoning capabilities. However, their inherent limitation to linguistic inputs constrains their flexibility in engaging with multimodal tasks. Recent advancements have sought to bridge this gap by enabling LLMs to process visual information, thereby extending their applicability to multimodal domains while preserving their intrinsic reasoning ability. To address this, recent efforts [25], [26], [27], [28], [60], [61] have led to the development of Multimodal Large Language Models (MLLMs) that integrate visual modalities into pre-trained LLM frameworks. These MLLMs typically adhere to a two-stage training approach. Initially, they employ a projection layer [28] or a Q-former [25], [27], [60] to align visual features with the LLM input space by training on expansive image-text pairs. To enhance MLLMs' instruction-following ability, certain models subsequently undergo further fine-tuning with instruction-following data. For instance, LLaVA constructs and fine-tunes

on the LLaVA-150 k dataset, a vision-language conversational corpus generated by GPT-4 [62], whereas MiniGPT4 [28] leverages ChatGPT-produced image captions, which are more extensive than its initial training data. InstructBLIP [27] distinguishes itself by using broader vision-language instruction data, encompassing both template-transformed and LLM-generated data. Despite these innovations, MLLMs essentially convert multimodal inputs into tokenised formats comprehensible to LLMs, a process that might omit vital information. In this paper, we propose a method that enriches MLLMs by transforming specific image regions related to the query into image captions, thereby providing supplementary information to augment the reasoning capabilities of MLLMs.

III. Problem Definition and Motivation

Let $\mathcal{D}_s = \{(v_i, q_i, a_i)\}_{i=1}^{N_s}$ be the N_s training samples of the VQA task, where (v_i, q_i, a_i) is the image, question, answer from the image set $\mathcal{V}$, question set $\mathcal{Q}$, and answer set $\mathcal{A}$, respectively. However, owing to the real-world complexity, the test samples $\mathcal{D}_t = \{(v_j, q_j)\}_{j=1}^{N_t}$ may have different distributions from the training data, making achieving excellent performance highly challenging. Fortunately, we may have a multimodal large language model M_l pre-trained on large image-text public data, which may remedy the issues of distribution differences. Without loss of generality, let M_s be a VQA specialised model [1], [2], [63], [64] trained on $\mathcal{D}_s$.

The Visual Question Answering (VQA) task aims to answer the textual questions in terms of the corresponding images. M_s transforms the VQA task into a multiple-class classification task, where the answer set $\mathcal{A}$ corresponds to the classes. Existing methods seek to overcome bias issues by introducing debiasing techniques during training time or test time. However, one may not access the OOD data until the model is evaluated. In this sense, introducing debiasing techniques at training time and choosing the best model on the basis of test data $\mathcal{D}_t$ (OOD) may be unavailable in the real world. Moreover, alleviating the bias issues at test time [19] requires updating the VQA models, which may result in catastrophic forgetting. One feasible solution to mitigate the above issues is to leverage the rich knowledge and excellent reasoning ability of the Large Language Models (LLMs) (e.g., Flan-T5 [20], Vicuna [21]). However, LLMs accept only the language modality as input, resulting in being stuck in a bottleneck when accomplishing multimodal tasks. To this end, we seek to adopt Multimodal Large Language Models (MLLMs) M_l (e.g., BLIP-2 [25], InstructBLIP [27]) to assist in achieving unbiased and reasonable reasoning, in which MLLMs can not only accept other input modalities in addition to language (e.g., images) but also leverage the rich knowledge and powerful reasoning ability of LLMs.

IV. Proposed Method

Adapting MLLMs M_l to downstream tasks (e.g., VQA) by fine-tuning is resource intensive and time-consuming, which conflicts with limited training data and computational resources in real-world deployment. Thus, in this paper, we explore leveraging the zero-shot reasoning capabilities of MLLMs to mitigate bias issues inherent in these tasks. Despite these advantages, employing MLLMs for reasoning across all samples is time-consuming and impractical for real-world applications. To address this challenge, we introduce an innovative VQA Collaborative Inference framework (VQA-CI) that combines a VQA specialist M_s with MLLMs M_l to facilitate complex reasoning tasks efficiently. Our VQA-CI initially involves processing the samples through M_s (e.g., UpDn [2]). Subsequently, based on a sample identification criterion applied to the predictions of M_s, we determine the necessity of further processing the samples with MLLMs M_l (e.g., BLIP-2 [25]). This decision is predicated on whether the samples are biased or unreliable. Note that despite the potential inaccuracies in the predictions of VQA specialists M_s, they contribute substantially to the MLLMs' reasoning process by providing valuable insights, such as candidate answers and identifying question-relevant image regions. The final predictions are derived from either the VQA specialists M_s or MLLMs M_l, depending on the outcome of this collaborative reasoning process. VQA-CI enables leveraging the strengths of MLLMs and VQA specialists, significantly enhancing the reasoning capabilities and facilitating deployment in a cloud-edge computing framework, thereby augmenting the practical applicability of VQA models. The overview and algorithm of our VQA-CI are shown in Fig. 3 and Algorithm 1, respectively.

Algorithm 1: Pipeline of VQA-CI

Require: Test samples $\mathcal{D}_t = \{(v_j, q_j)\}_{j=1}^{N_t}$, the pre-trained VQA specialist M_s, the Multimodal Large Language Model M_l, batch size B.

1: **for** a mini-batch $\mathcal{D}_b = \{(v_b, q_b)\}_{b=1}^{B}$ in $\mathcal{D}_t$ **do**
2: Generate the negative samples $(\bar{v}_j, q_b)$ and $(v_b, \bar{q}_j)$ for each sample (v_b, q_b) via randomly sampling images or questions in $\mathcal{D}_b$.
3: Calculate the predictions $\mathbf{p}$, $\bar{\mathbf{p}}_v$ and $\bar{\mathbf{p}}_q$ for the samples (v_b, q_b) and the counterpart negative samples $(\bar{v}_j, q_b)$ and $(v_b, \bar{q}_j)$ with M_s via Eq. (1).
4: Based on $\mathbf{p}$, $\bar{\mathbf{p}}_v$ and $\bar{\mathbf{p}}_q$, obtain the valid samples indicator $\mathbf{I}$ via Eq. (4).
5: *// The samples are either biased or unreliable*
6: **if** $\mathbf{I}$ is False **then**
7: Obtain the Top-k candidate answers $\mathcal{A}_{sub}$, and question-relevant image regions $\mathcal{V}_{sub}$ using the image attention from M_s.
8: Adopt PromptCap [31] to generate captions c on the identified question-relevant image regions $\mathcal{V}_{sub}$.
9: Obtain final prediction $\mathbf{p}$ by forwarding M_l with the prompt that contains instruction, sample (v_b, q_b), captions c, and candidate answers $\mathcal{A}_{sub}$.
10: **end if**
11: **end for**

Ensure: The predictions $\{\mathbf{p}_j\}_{j=1}^{N_t}$ for all $(v_j, q_j) \in \mathcal{D}_t$.

A. Reasoning With VQA Specialists

We devise two principal instance-level indicators to determine whether to accept the preliminary outputs of M_s or to defer

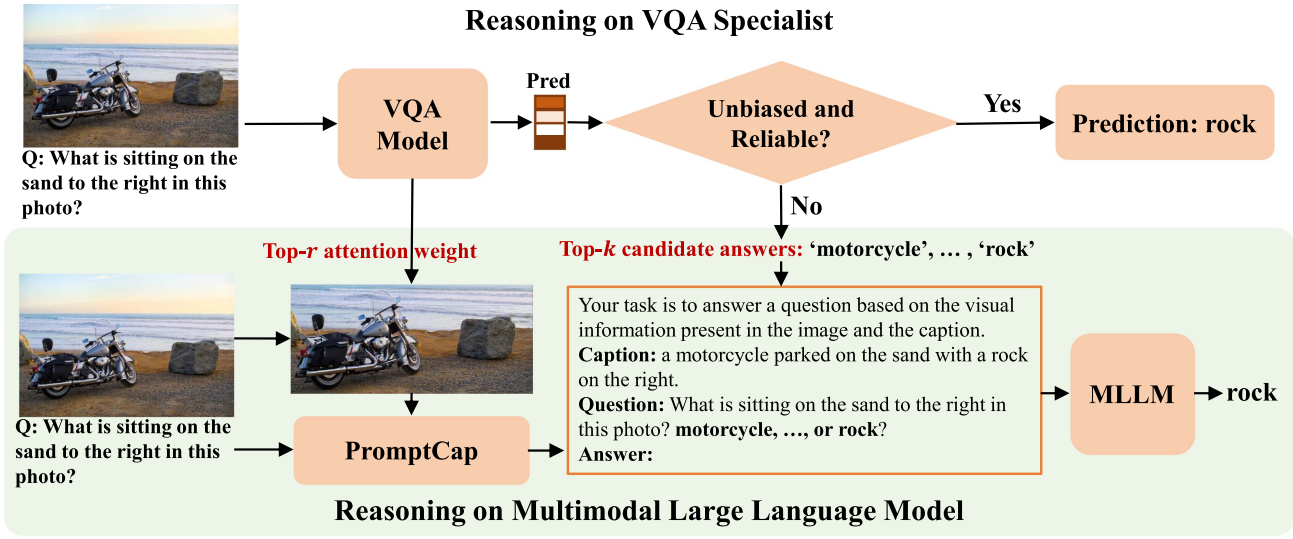

Fig. 3. Overview of our VQA-CI. The test samples are first forwarded with VQA specialists, and the predictions are accepted if the samples are deemed unbiased and reliable. Otherwise, the samples require re-evaluation by the MLLMs. Before processing the samples with MLLMs, we extract discriminative information from VQA specialists, i.e., question-related image captions and top-k candidate answers, to form prompts with rich information for assisting the reasoning of MLLMs.

decision-making to M_l for more robust inference (i.e., biased vs. unbiased and reliable vs. unreliable).

Instance-level Indicators: Inspired by [16], [19], [65], our VQA-CI encompasses two critical considerations: 1) whether M_s uses the captured biases inherent in the training set to answer the questions, and 2) whether M_s makes predictions with high confidence. Specifically, the first aspect emphasises the importance of leveraging reasoning over reliance on dataset biases for answering questions, whereas the second aspect underscores the necessity for predictions to be made with a high degree of certainty, thereby indicating their reliability.

1) Unbiased Sample Identification: Bias issues in VQA often indicate a tendency of VQA models to rely on unimodal information to answer questions related to images [14], [16], [32]. From this perspective, assume that given a pair of biased test samples (v_i, q_i), VQA specialists M_s may generate similar predictions regardless of which image or question it is, i.e., $(\bar{v}_j, q_i)$ or $(v_i, \bar{q}_j)$.

Based on this intuition, one straightforward way to identify biased samples involves the construction of negative samples to induce VQA specialists M_s to generate predictions similar to those of the original samples. Specifically, given a mini-batch of test samples $\{(v_i, q_i)\}_{i=1}^{B}$, for each sample (v_i, q_i), we first randomly sample one question $\bar{q}_j$ from $\{q_i\}_{i=1}^{B}$ and one image $\bar{v}_j$ from $\{v_i\}_{i=1}^{B}$. These sampled elements (i.e., $\bar{q}_j$ and $\bar{v}_j$) are then paired with the original image v_i and question q_i to construct two types of negative samples: $(\bar{q}_j, v_i)$ and $(q_i, \bar{v}_j)$. Subsequently, we feed the original samples (v_i, q_i) along with two types of negative samples $(\bar{q}_j, v_i)$ and $(q_i, \bar{v}_j)$ into M_s to obtain the corresponding predictions $\mathbf{p}$, $\bar{\mathbf{p}}_v$ and $\bar{\mathbf{p}}_q$ which can be formulated as:

$$\mathbf{p} = M_s(\mathcal{A}|v_i, q_i), \bar{\mathbf{p}}_v = M_s(\mathcal{A}|\bar{v}_j, q_i), \bar{\mathbf{p}}_q = M_s(\mathcal{A}|v_i, \bar{q}_j). \tag{1}$$

Then, we classify the samples (v_i, q_i) as biased if the answer derived from $\mathbf{p}$ coincides with that obtained from either $\bar{\mathbf{p}}_v$ or $\bar{\mathbf{p}}_q$. Hence, the biased indicator function is defined as:

$$\begin{aligned}\mathbf{I}_b(v_i, q_i) = \; & \mathbb{I}_{\{\text{argmax}(\mathbf{p}) \neq \text{argmax}(\bar{\mathbf{p}}_v)\}}(v_i, q_i) \cdot \\ & \mathbb{I}_{\{\text{argmax}(\mathbf{p}) \neq \text{argmax}(\bar{\mathbf{p}}_q)\}}(v_i, q_i),\end{aligned} \tag{2}$$

where the operation argmax selects the index corresponding to the maximum value within the predictions. Here, $\mathbb{I}(\cdot)$ denotes an indicator function, i.e., returns a value of 1 if argmax$(\mathbf{p})$ does not equal argmax$(\bar{\mathbf{p}}_v)$ and a value of 0 otherwise.

Note that in certain exceptional scenarios, where only a single sample is available in each mini-batch, i.e., $B = 1$, the construction of negative samples through sampling from within the mini-batch is not feasible. To address this limitation, we propose the generation of Gaussian noise with the same dimensions as the question or image features to substitute the negative part. We present the experimental results regarding $B = 1$ in Tables IX and X.

2) Reliable Sample Identification: In addition to identifying unbiased samples, evaluating the reliability of predictions rendered by M_s is imperative to ascertain whether they can be confidently accepted as the final answer. To this end, we adopt entropy as a formal measure of prediction reliability. Specifically, a high entropy value indicates a model's uncertainty regarding a sample, deeming the prediction unreliable, and vice versa. Given the predictions $\mathbf{p} = M_s(\mathcal{A}|v_i, q_i)$, we calculate the entropy of the samples as $E(v_i, q_i) = -\sum \mathbf{p} \log \mathbf{p}$. To recognise the reliable samples, we introduce a pre-defined entropy threshold E_t. Samples whose entropy is lower than E_t are considered reliable. The criterion for determining reliability is thus formulated as follows:

$$\mathbf{I}_r(v_i, q_i) = \mathbb{I}_{\{E(v_i, q_i) < E_t\}}(v_i, q_i). \tag{3}$$

In this paper, the entropy threshold E_t is determined by the number of candidate answers C, that is, $E_t = \alpha \ln C$, where α is a hyper-parameter within the range $[0, 1]$. In this way, we can dynamically select reliable samples to adapt to the computational and time limitations of the real-world scenario by adjusting the entropy threshold E_t.

Leveraging the aforementioned criteria, it becomes feasible to identify valid samples that are both reliable and unbiased with respect to M_s, i.e.,

$$\mathbf{I}(v_i, q_i) = \mathbf{I}_r(v_i, q_i) \cdot \mathbf{I}_b(v_i, q_i). \tag{4}$$

B. Reasoning With Multimodal Large Language Models

For the remaining samples, which are either unreliable or biased, M_s is incapable of delivering reliable predictions. Thus, we seek to adopt MLLMs M_l to generate more convincing answers. In scenarios involving out-of-distribution test samples, VQA specialists M_s often exhibit sub-optimal performance. Intriguingly, as illustrated in Fig. 2, despite the UpDn model registering a modest top-1 accuracy of 41%, it remarkably achieves over 80% in top-5 accuracy. This phenomenon demonstrates that although VQA models suffer from bias issues, they still contain some discriminative knowledge. Thus, before assigning the invalid samples to MLLMs, it is beneficial to extract valuable insights from VQA specialists to enhance the prediction capabilities of MLLMs.

1) Knowledge Mining in VQA Specialists: As shown in Fig. 2, the VQA specialists (e.g., UpDn [2]) have a high top-5 accuracy. This suggests that VQA specialists still possess the ability to identify question-related image regions to answer questions. In light of this, we propose extracting two types of information from VQA specialists: 1) candidate answers, and 2) question-related image regions. The candidate answers can help MLLMs streamline the reasoning process, i.e., regarding the VQA task as an answer re-ranking task. Moreover, identifying question-related image regions enables the provision of detailed information for reasoning, such as by employing image captioning models to generate fine-grained captions associated with the questions. Our method not only enhances the reasoning capabilities of MLLMs but also leverages the discriminative knowledge embedded within VQA specialists.

Candidate Answers: When MLLMs are adopted to conduct VQA tasks, these models require generating textual responses from an extensive vocabulary, presenting a significant challenge. To mitigate this difficulty, we propose transforming the generation task into an answer re-ranking process, thereby simplifying the VQA task. Specifically, for each instance, we extract the top-k answers $\mathcal{A}_{\text{sub}} = \{a_i\}_{i=1}^{k}$ from the predictions made by the VQA specialists to serve as candidate answers. A textual prompt incorporating these candidates is subsequently constructed to compel the MLLMs to select the most appropriate answer from among these candidates.

Question-Related Image Regions: MLLMs are capable of performing tasks across different modalities by transforming information from one modality (e.g., images) into a format comprehensible by LLMs. However, the transformation process may be sub-optimal. MLLMs often summarise the information from images into a limited number of token features (e.g., Q-Former [25]), which may result in information loss. To alleviate this issue, inspired by [29], our method involves enriching MLLMs with additional textual information derived from images to increase performance. Nevertheless, as shown by [29], equipping general captions for augmenting image information may not always lead to performance improvements. This can lead to worse outcomes than scenarios where no captions are introduced. This phenomenon suggests that textual information, which is unrelated to the question, may introduce noise and thereby degrade model performance.

To this end, we identify question-related image regions by leveraging visual attention from VQA specialists. Specifically, given a sample (v_i, q_i), we first feed it to the VQA specialist M_s and obtain the visual attention weights related to the image v_i. Subsequently, we select image regions corresponding to the top-r visual attention weights as candidate question-related image regions $\mathcal{V}_{sub} = \{\tilde{v}_j\}_{j=1}^{r}$, where r serves as the hyperparameter.

Question-relevant Caption Generation: On the basis of the identified question-relevant image regions, we adopt image caption models [25], [31] to generate captions that are pertinent to the posed questions. In general, the question-relevant image regions typically concentrate on specific segments of the image. Consequently, generating captions for all regions in $\mathcal{V}_{sub}$ is unnecessary. To this end, we propose amalgamating regions $\mathcal{V}_{sub} = \{\tilde{v}_j\}_{j=1}^{r}$ into one target question-related region. Subsequent to this consolidation, since PromptCap [31] is well-trained to generate question-related captions, we employ PromptCap to serve as our caption generator. The prompt of caption generation is: "*Please describe this image according to the given question:*". Through this procedure, we are able to procure question-relevant captions c for each sample.

2) Prompt Design: To construct complete prompts for MLLMs, we employ a structured approach that concatenates instructions, captions, questions, and candidate answers. The instruction text is articulated as follows: "Your task is to answer a question based on the visual information present in the image and the caption.". The caption is introduced in the following format: "Caption: $[c]$". For the question and answer segment, the format is specified as "Question: $[q][\mathcal{A}_{sub}]$ Short Answer:", where $\mathcal{A}_{sub}$ represents the subset of candidate answers. Following [29], we concatenate all candidate answers subsequent to the question, exemplified by "What colour of the dishes? yellow, orange, $\cdots$, or red? Short Answer:". We subsequently adopt a beam search strategy on the MLLMs to derive the answer.

V. Experiments

A. Experimental Setup

1) Models: We employ UpDn [2] pre-trained on the training set of the VQA-CP v2 or GQA-OOD datasets as a VQA specialist M_s. For the Multimodal Large Language Models (MLLMs) M_l, we primarily examine two types of MLLMs: i.e., BLIP-2 [25] (BLIP-2 Flan T5 XL and BLIP-2 OPT 6.7B) and instructBLIP [27] (InstructBLIP Flan T5 XL and InstructBLIP

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE GQA-OOD VALIDATION AND TEST SETS IN TERMS OF ACCURACY (%)

	Model	validation set (%)			testdev set (%)		
		All	Head	Tail	All	Head	Tail
Baseline	Quest. Prior	-	-	-	21.60	24.10	17.80
	LSTM [3]	-	-	-	30.70	34.80	24.00
	UpDn† [2]	52.55	58.87	40.09	43.38	46.51	38.76
Training Time	BP [66]	-	-	-	33.10	34.50	30.80
	LM [66]	-	-	-	34.50	35.90	32.20
	RUBI [32]	-	-	-	38.80	40.80	35.70
	CSS [43]	-	-	-	44.24	41.20	46.11
	RUBi+QB	-	-	-	46.70	42.10	**49.40**
Test Time	TDS† [19]	52.14	58.11	40.34	43.13	45.59	39.13
	Tent† [53]	52.33	59.16	38.84	43.53	45.70	39.98
	ETA† [51]	52.51	59.00	39.70	43.60	45.93	39.79
MLLMs	BLIP-2 (OPT) [25]	37.59	40.33	32.18	31.80	32.31	30.95
	BLIP-2 (Flan T5) [25]	45.57	49.60	37.60	38.63	40.68	35.28
	InstructBLIP (Flan T5) [27]	50.51	54.78	42.09	44.67	46.28	42.05
	InstructBLIP (Vicuna) [27]	52.11	56.81	42.83	43.96	44.78	42.62
Ours	BLIP-2 (OPT) + VQA-CI	48.85	53.85	39.00	40.52	42.47	37.35
	BLIP-2 (Flan T5) + VQA-CI	57.43	62.51	**47.38**	47.28	49.45	43.74
	InstructBLIP (Flan T5) + VQA-CI	56.69	61.43	47.31	47.96	50.49	43.84
	InstructBLIP (Vicuna) + VQA-CI	**58.18**	**63.82**	47.04	**48.10**	**51.01**	43.37

The backbone of all VQA specialists is UpDn [2]. Overall best scores are **bold**. † denotes our re-implementation of the methods.

Vicuna 7B). All mentioned MLLMs are accessible through the HuggingFace platform.

2) Datasets: We conduct experiments to evaluate our VQA-CI on two benchmark out-of-distribution (OOD) datasets: GQA-OOD [30] and VQA-CP v2 [15].

The **GQA-OOD dataset** is derived from the GQA [67] dataset, where it shares the same training set with GQA while re-organising the validation and test sets to introduce distribution shifts. The training set of GQA-OOD contains approximately 72 k images and 943 k questions, the validation set includes approximately 9.4 k images and 51 k questions, and the test set contains 388 images and 2.8 k questions.

The **VQA-CP v2 dataset** is constructed by re-organising the training and validation sets of the VQA v2 [12] dataset. The training set of VQA-CP v2 contains approximately 121 k images and 483 k questions, whereas the test set contains approximately 98 k images and 220 k questions. Note that InstructBLIP-based models have been pre-trained on the training and validation sets of the VQA v2 dataset, which may affect the outcomes when applied to the VQA-CP v2 dataset. Therefore, to ensure a more robust evaluation and mitigate any potential biases introduced during the pre-training phase, we conduct our experiments on the VQA-CP v2 dataset using BLIP-2-based models.

3) Implementation Details: When extracting discriminative information from VQA specialists, we consider two hyperparameters: the number of candidate answers k and the number of question-related image regions r. For candidate answers, we set $k = 5$ for both the GQA-OOD and VQA-CP v2 datasets. For question-related image regions r, the datasets vary in their image region features: the VQA-CP v2 dataset uses the top 36 object features per image extracted by Faster R-CNN [68], whereas the GQA-OOD dataset includes up to 100 image region features per image. This discrepancy makes it impractical to set a fixed number of image regions across datasets. To address this, we define r as a ratio, setting r to 0.8 and 0.5 for the GQA-OOD and VQA-CP v2 datasets, respectively. For the generation process of MLLMs, we adopt a beam search strategy with the number of beams set to 5. The source codes are available at https://github.com/Zhiquan-Wen/VQA-CI.

B. Quantitative Results

We compare our VQA-CI with other state-of-the-art methods on the GQA-OOD [30] and VQA-CP v2 [15] datasets, and report the experimental results in Tables I and II, respectively. From the results presented in Table I regarding the GQA-OOD dataset, we have the following observations: 1) Although the training-time methods adopt OOD validation sets to select models, they achieve comparable or even inferior performance to the UpDn models. This outcome suggests that traditional debiasing techniques [32], [43], [66] may lack generalisability, thereby limiting their applicability in real-world scenarios. 2) Test-time methods exhibit performance on par with the UpDn model. Notably, on the testdev set, these methods perform better on "Tail" data (OOD) but worse on "Head" data (IID) than UpDn does. This finding indicates that test time methods may suffer from the issue of catastrophic forgetting. 3) The zero-shot results of MLLMs are comparable to those of UpDn, highlighting the importance and necessity of leveraging the exceptional zero-shot reasoning capabilities of MLLMs to further enhance performance. 4) Integrating VQA-CI with MLLMs leads to a significant improvement in performance. Specifically, when BLIP-2 models (either OPT or Flan T5) are equipped with VQA-CI, they

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE VQA-CP V2 TEST SET IN TERMS OF ACCURACY (%)

	Model	VQA-CP v2 test set(%)			
		All	Yes/No	Num	Other
Baseline	SAN [69]	24.96	38.35	11.14	21.74
	GVQA [15]	31.30	57.99	13.68	22.14
	UpDn [2]	40.94	43.87	12.90	47.09
	UpDn† [2]	41.52	43.03	13.53	48.41
Training Time	CF-VQA [18]	53.55	**91.15**	13.03	44.97
	SSL-VQA [17]	57.59	86.53	29.87	50.03
	CSS [43]	58.95	84.37	49.42	48.21
	CSS+SBS [37]	59.57	87.44	52.96	46.79
	CVIV+iter [39]	60.08	88.85	40.77	50.30
	PW-VQA [36]	60.26	88.09	**59.13**	45.99
	DDG [46]	61.14	88.77	49.33	49.90
	Mutant [42]	61.72	88.90	49.68	50.78
	D-VQA [16]	61.91	88.93	52.32	50.39
Test Time	Tent [53]	41.17	43.88	13.95	47.22
	ETA [51]	41.28	43.81	13.76	47.51
	LAME [52]	41.37	42.91	15.70	47.61
	TDS [19]	48.28	66.23	16.74	47.53
MLLMs	BLIP-2 (OPT) [25]	51.17	75.88	22.72	46.02
	BLIP-2 (Flan T5) [25]	59.63	84.29	46.21	50.39
Ours	BLIP-2 (OPT) + VQA-CI	54.14	71.13	31.39	51.49
	BLIP-2 (Flan T5) + VQA-CI	**67.22**	84.84	<u>56.88</u>	**60.83**

The backbone of all VQA specialists is UpDn [2]. Overall best scores are **bold**, and the second best of our overall scores are <u>underlined</u>. † denotes our re-implementation of the methods.

outperform configurations without VQA-CI by approximately 11% and 9% on the validation and testdev sets, respectively.

From the results in Table II regarding the VQA-CP v2 dataset, several key observations emerge: 1) The training-time methods outperform the UpDn model by a considerable margin (e.g., D-VQA [16] surpasses UpDn by approximately 20%). However, the reliance on an OOD test set for selecting optimal models limits their practical application. 2) The test-time methods also outperform the UpDn model, and the extent of improvement is not as pronounced as that of the training-time methods. Nevertheless, test-time methods update models in the absence of ground-truth answers and direct access to original data, potentially leading to catastrophic forgetting. 3) Zero-shot MLLMs demonstrate superior performance. Incorporating our VQA-CI further enhances model performance. Specifically, the BLIP-2 (Flan T5) model equipped with VQA-CI outperforms those without it by approximately 8%, underscoring the effectiveness of VQA-CI. These findings underscore the effectiveness of our VQA-CI.

C. Qualitative Results

To further highlight the effectiveness of our VQA-CI, we present visualisation results on the GQA-OOD validation set in Fig. 4. From the results, we display the captions generated by PromptCap [31] and VQA-CI. The results indicate PromptCap may generate general captions or even wrong information, potentially misleading the reasoning processes of MLLMs (e.g., BLIP-2 Flan T5). In contrast, VQA-CI focuses on generating captions for image regions that are related to the posed questions, resulting in more accurate and relevant captions. This approach significantly enhances the relevance and accuracy of the information provided to MLLMs, thereby demonstrating the superiority of our VQA-CI.

TABLE III
ABLATION STUDIES ON THE IMPACT OF VARYING CANDIDATE ANSWERS ON THE GQA-OOD DATASET IN TERMS OF ACCURACY (%)

Model	validation set (%)			testdev set (%)		
	All	Head	Tail	All	Head	Tail
BLIP-2 (Flan T5) [25]	45.57	49.60	37.60	38.63	40.68	35.28
+ VQA-CI (top-3 answers)	57.32	62.38	47.33	47.28	49.86	43.09
+ VQA-CI (top-5 answers)	57.43	62.51	47.38	47.28	49.45	43.74
+ VQA-CI (top-7 answers)	57.47	62.51	47.50	47.07	49.62	42.90
+ VQA-CI (top-8 answers)	57.39	62.45	47.39	46.85	49.22	42.99

D. Ablation Studies

1) Effect of the Number of Candidate Answers $\mathcal{A}_{sub}$: $\mathcal{A}_{sub}$ denotes available candidate answers in the prompt that MLLMs can select in the reasoning process. To evaluate the effect of $\mathcal{A}_{sub}$, we conduct experiments on the GQA-OOD with BLIP-2 Flan T5 as MLLMs and report the results in Table III. The results indicate that 1) as the number of candidate answers increases (i.e., from the top-3 answers to the top-5 answers), the model performance gradually improves. These results demonstrate that a broader set of candidate answers increases the chances of including the correct target answer, thereby facilitating more effective reasoning by the MLLMs. 2) A diminishing return in performance occurs when the number of candidate answers exceeds five. This decline suggests that an overload of candidate answers may introduce noise, potentially impeding the models' reasoning processes. We thus select the top-5 candidate answers in all the experiments.

2) Evaluation of the Question-Related Regions r: r signifies the selection of the top-r regions based on the image attention weights from the VQA specialists (e.g., UpDn), which are then aggregated to form a sub-image as question-related image regions. To accommodate the variability in the number

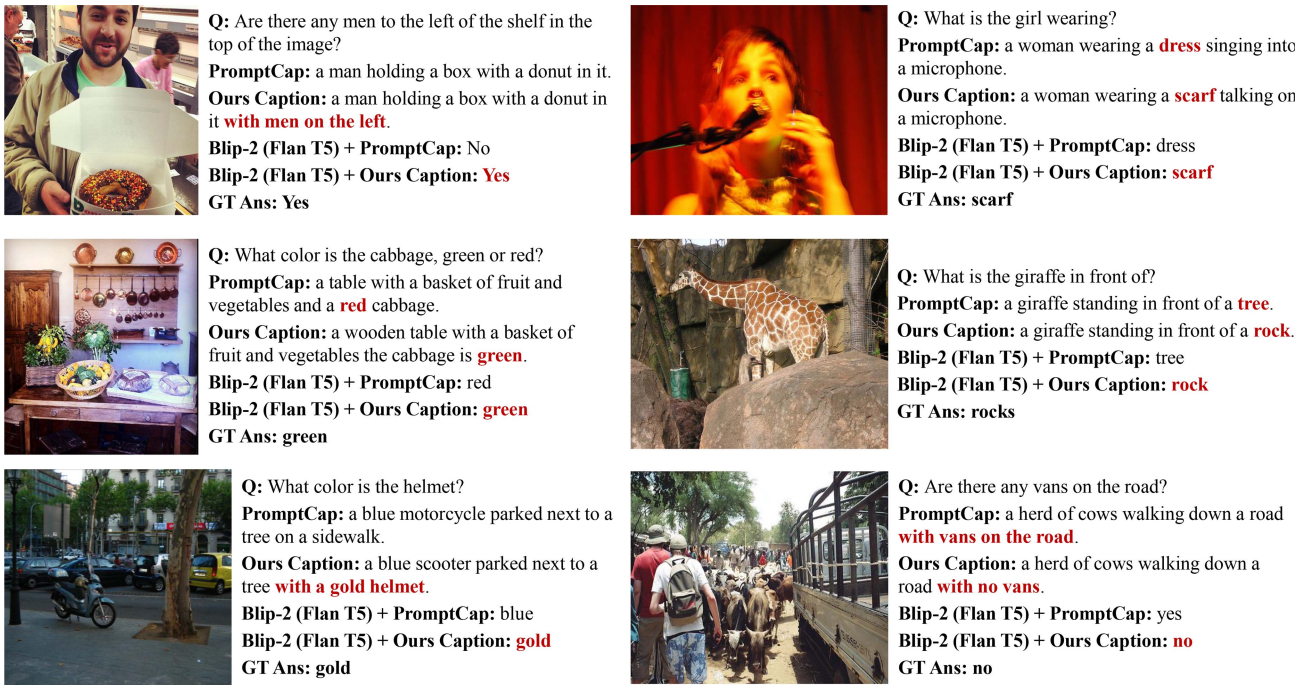

Fig. 4. Qualitative comparison among the captions generated by PromptCap [31] and our VQA-CI on the GQA-OOD validation set.

TABLE IV
ABLATION STUDIES ON THE EFFECT OF DIFFERENT VQA SPECIALISTS ON THE VQA-CP V2 DATASET

Model	VQA Specialists	VQA-CP v2 test set (%)			
		All	Yes/No	Num	Other
UpDn† [2]	-	41.52	43.03	13.53	48.41
LXMERT† (w/o pre-trained) [70]	-	41.72	43.61	14.17	48.30
LXMERT† [70]	-	52.16	49.06	24.23	61.44
BLIP-2 (Flan T5) + VQA-CI	UpDn	67.22	84.84	56.88	60.83
BLIP-2 (Flan T5) + VQA-CI	LXMERT	**68.64**	**85.18**	**57.72**	**62.97**

The bold values indicate the best performance scores.

TABLE V
ABLATION STUDIES ON THE EFFECT OF DIFFERENT SCALES OF MLLMS ON THE GQA-OOD AND VQA-CP V2 DATASETS

Model	GQA-OOD testdev set (%)			VQA-CP v2 test set (%)			
	All	Head	Tail	All	Yes/No	Num	Other
UpDn† [2]	43.38	46.51	38.76	41.52	43.03	13.53	48.41
BLIP-2 (Flan T5) [25]	38.63	40.68	35.28	59.63	84.29	46.21	50.39
BLIP-2 (Flan T5 XXL) [25]	42.10	44.32	38.48	59.62	85.24	32.49	53.64
BLIP-2 (Flan T5) + VQA-CI	47.28	49.45	**43.74**	67.22	84.84	56.88	**60.83**
BLIP-2 (Flan T5 XXL) + VQA-CI	**48.96**	**52.28**	43.56	**68.12**	**87.27**	**59.51**	60.46

The bold values indicate the best performance scores.

of image regions across different images (e.g., the GQA-OOD dataset), we normalise r into a ratio, i.e., selecting the top-$\lceil r * x \rceil$ regions for an image with x regions. To evaluate r, we conduct experiments on the GQA-OOD and VQA-CP v2 datasets with different r values and present the experimental results in Tables VI and VII, respectively. From the results on the GQA-OOD dataset in Table VI, the optimal performance on the GQA-OOD dataset is achieved with $r = 0.8$, with any deviation from this value resulting in diminished accuracy. These results indicate that a small r may lead to the loss of critical regions, whereas a high r may introduce background information that may affect question-relevant caption generation. We adopt $r = 0.8$ on the GQA-OOD dataset for all the experiments. Similar results are shown for the VQA-CP v2 dataset in Table VII, we then adopt $r = 0.5$ on the VQA-CP v2 dataset for all the experiments.

3) Effect of Different VQA Specialists: To investigate the impact of various VQA specialists, we conducted ablation studies

TABLE VI
ABLATION STUDIES ON THE EFFECT OF THE NUMBER OF QUESTION-RELATED REGIONS IN TERMS OF ACCURACY (%)

Model	GQA-OOD testdev set (%)		
	All	Head	Tail
BLIP-2 (Flan T5) [25]	38.63	40.68	35.28
+ VQA-CI (r=0.5)	46.82	49.22	42.90
+ VQA-CI (r=0.7)	47.17	49.51	43.37
+ VQA-CI (r=0.8)	**47.28**	49.45	43.74
+ VQA-CI (r=0.9)	47.10	49.28	43.56

The bold values indicate the best performance scores.

TABLE VII
ABLATION STUDIES ON THE EFFECT OF THE NUMBER OF QUESTION-RELATED REGIONS IN TERMS OF ACCURACY (%)

Model	VQA-CP v2 test set (%)			
	All	Yes/No	Num	Other
BLIP-2 (Flan T5) [25]	59.63	84.29	46.21	50.39
+ VQA-CI (r=0.3)	67.20	84.87	56.88	60.77
+ VQA-CI (r=0.5)	67.22	84.84	56.88	60.83
+ VQA-CI (r=0.7)	67.23	84.88	56.91	60.81
+ VQA-CI (r=0.8)	67.23	84.87	56.92	60.81

on the VQA-CP v2 dataset using BLIP-2 (Flan T5) as MLLMs. Moreover, UpDn [2] and LXMERT [70] are selected as the VQA specialists. The results presented in Table IV yield the following insights: 1) Pre-trained LXMERT exhibits more robust performance than LXMERT without pre-training does, suggesting that pre-training is an available way to alleviate bias issues, but its performance is still unsatisfactory. 2) LXMERT + VQA-CI outperforms UpDn + VQA-CI by approximately 1%. These findings indicate that more advanced VQA specialists can provide more accurate question-related regions and candidate answers, thereby enhancing the reasoning capabilities of MLLMs. Furthermore, the above results demonstrate that our VQA-CI is a VQA specialist-agnostic approach, underscoring its general applicability in real-world scenarios.

4) Impact of the MLLM Scale on Model Performance: To evaluate the impact of the MLLM scale on model performance, we conduct experiments on the GQA-OOD testdev set and the VQA-CP v2 test set using two BLIP-2 variants: Flan T5 XL and Flan T5 XXL. From the results in Table V, we have the following observations: 1) BLIP-2 Flan T5 XXL outperforms BLIP-2 Flan T5 XL, with a notable 3.5% improvement on the GQA-OOD dataset, indicating that larger MLLMs possess stronger reasoning abilities, although overall performance remains suboptimal. 2) When equipped with VQA-CI, BLIP-2 Flan T5 XXL outperforms BLIP-2 Flan T5 XL by approximately 1% on both datasets, demonstrating that larger MLLMs can more effectively leverage the supplementary information from VQA-CI for enhanced reasoning. 3) Regardless of the scale of MLLMs, integrating VQA-CI significantly boosts model performance, highlighting the effectiveness of our VQA-CI.

5) Effect of Each Component on Model Performance: To assess the impact of distinct components in our VQA-CI, we conduct ablation studies on the GQA-OOD dataset across different MLLMs. From these results in Table VIII, we draw several key observations: 1) Integration of MLLMs with PromptCap [31] significantly enhances model performance. (e.g., on the validation set, BLIP-2 (Flan T5) + PromptCap outperforms that without PromptCap by approximately 7%.). These results indicate that MLLMs may be dominated by LLMs, and transforming other modalities into tokens may introduce information loss. 2) Using the question-related image regions to generate captions c further improves the model performance (e.g., on the testdev set, InstructBLIP (Vicuna) combined with captions c outperforms the same configuration enhanced with PromptCap by 1%). These results illustrate the importance of recognising the target image regions to generate supplemental textual information to assist the reasoning of MLLMs. 3) Incorporating candidate answers $\mathcal{A}_{sub}$ into MLLMs significantly improves model performance. For example, on the testdev set, BLIP-2 (Flan T5) + captions c + $\mathcal{A}_{sub}$ outperforms BLIP-2 (Flan T5) + captions c by approximately 5%. This finding demonstrates that providing candidate answers can reduce the reasoning complexity for MLLMs, leading to improved model performance. Overall, the above results demonstrate the effectiveness of each component in our VQA-CI.

TABLE VIII
ABLATION STUDIES ON THE EFFECT OF EACH COMPONENT ON THE GQA-OOD DATASET IN TERMS OF ACCURACY (%)

Model	validation set (%)			testdev set (%)		
	All	Head	Tail	All	Head	Tail
BLIP-2 (OPT) [25]	37.59	40.33	32.18	31.80	32.31	30.95
+ PromptCap [31]	43.77	46.98	37.43	36.66	37.39	35.47
+ Captions c	43.71	46.98	37.26	37.66	38.20	36.78
+ Captions c + $\mathcal{A}_{sub}$	48.85	53.85	39.00	40.52	42.47	37.35
BLIP-2 (Flan T5) [25]	45.57	49.60	37.60	38.63	40.68	35.28
+ PromptCap [31]	52.11	56.13	44.19	42.35	44.78	38.38
+ Captions c	53.03	57.16	44.88	42.81	45.07	39.13
+ Captions c + $\mathcal{A}_{sub}$	57.43	62.51	47.38	47.28	49.45	43.74
InstructBLIP (Flan T5) [27]	50.51	54.78	42.09	44.67	46.28	42.05
+ PromptCap [31]	53.35	57.43	45.31	45.10	46.97	42.05
+ Captions c	53.58	57.67	45.50	45.35	47.09	42.52
+ Captions c + $\mathcal{A}_{sub}$	56.69	61.43	47.31	47.96	50.49	43.84
InstructBLIP (Vicuna) [27]	52.11	56.81	42.83	43.96	44.78	42.62
+ PromptCap [31]	53.68	57.93	45.28	43.03	43.80	41.77
+ Captions c	53.83	57.93	45.71	44.13	44.95	42.80
+ Captions c + $\mathcal{A}_{sub}$	58.18	63.82	47.04	48.10	51.01	43.37

6) Effect of Collaborative Inference: To demonstrate the effectiveness of collaborative inference by integrating MLLMs and VQA specialists, we conduct ablation studies on the GQA-OOD and VQA-CP v2 datasets and report the results in Tables IX and X, respectively. From the results on the GQA-OOD dataset in Table IX, we obtain several key insights: 1) A progressive enhancement in collaborative inference performance is observed with the reduction in the entropy threshold α, indicating that as samples become more reliable, the likelihood of accurate predictions by the VQA specialist increases. This trend underscores the significance of sample reliability in the inference process. 2) Only $\mathbf{I}_b$ and only $\mathbf{I}_r$ ($\alpha = 0.1$) outperform the baseline UpDn, which demonstrates the effectiveness of our biased and unreliable sample indicators. Note that $\mathbf{I}_r$ demonstrates superior performance compared with $\mathbf{I}_b$, suggesting that issues of unreliability may have a more pronounced impact than biases do. The composition of the GQA-OOD dataset supports this conclusion, with the validation set comprising 33,882 samples in the "Head"

TABLE IX
ABLATION STUDIES ON THE EFFECT OF COLLABORATIVE INFERENCE BY INTEGRATING MLLMS AND UPDN ON THE GQA-OOD DATASET IN TERMS OF ACCURACY (%)

Model	validation set (%)			# Samples	testdev set (%)			# Samples
	All	Head	Tail		All	Head	Tail	
UpDn [2]	52.55	58.87	40.09	-	43.38	46.51	38.76	-
+$\mathbf{I}_r$ ($\alpha = 0.1$)	57.78	64.11	45.30	25,167 (49.3%)	46.39	48.53	42.90	1,459 (52.1%)
+$\mathbf{I}_b$	53.14	58.38	42.79	19,927 (39.0%)	45.06	48.99	38.66	1,062 (37.9%)
+$\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.5$)	53.27	58.46	43.03	20,532 (40.2%)	45.17	49.05	38.85	1,118 (39.9%)
+$\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.3$)	54.41	59.53	44.29	24,975 (48.9%)	45.60	49.45	39.32	1,369 (48.9%)
+$\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.2$)	56.29	61.47	46.08	32,213 (63.1%)	46.03	49.22	40.83	1,746 (62.4%)
+$\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.1$)	57.25	62.56	46.78	39,817 (78.0%)	46.92	49.74	42.33	2,177 (73.1%)
Ours (BLIP-2 Flan T5)	57.43	62.51	47.38	51,045 (100%)	47.28	49.45	43.74	2,796 (100%)
Mini-Batch Size is 1								
UpDn† [2]	52.55	58.87	40.09	-	43.38	46.51	38.76	-
+ $\mathbf{I}_r$ ($\alpha = 0.1$)	57.78	64.11	45.30	25,167 (49.3%)	46.39	48.53	42.90	1,459 (52.1%)
+ $\mathbf{I}_b$	52.67	58.93	40.31	2,989 (5.8%)	43.35	46.28	38.57	167 (5.9%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.5$)	52.81	59.01	40.58	3,616 (7.0%)	43.49	46.34	38.85	228 (8.1%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.3$)	54.06	60.18	41.97	8,552 (16.7%)	43.99	46.86	39.32	496 (17.7%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.2$)	56.43	62.57	44.31	17,302 (33.8%)	44.74	46.91	41.20	964 (34.4%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.1$)	57.70	64.01	45.25	26,581 (52.0%)	45.99	47.89	42.90	1,526 (54.5%)
Ours (BLIP-2 Flan T5)	57.43	62.51	47.38	51,045 (100%)	47.28	49.45	43.74	2,796 (100%)

The MLLMs are BLIP-2 Flan T5. "# samples" denotes the number of samples MLLMs will forward. α is a hyper-parameter that adjusts the entropy threshold. I_b is a biased sample indicator, whereas I_r is an unreliable sample indicator.

TABLE X
ABLATION STUDIES ON THE EFFECT OF COLLABORATIVE INFERENCE BY INTEGRATING MLLMS AND VQA SPECIALISTS ON THE VQA-CP V2 DATASET

Model	VQA-CP v2 test set(%)				# Samples
	All	Yes/No	Num	Other	
UpDn† [2]	41.52	43.03	13.53	48.41	-
+ $\mathbf{I}_r$ ($\alpha = 0.1$)	49.84	43.61	28.37	59.01	107,569 (48.9%)
+ $\mathbf{I}_b$	60.20	83.85	45.37	51.87	103,625 (47.1%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.5$)	60.60	83.97	45.51	52.50	115,106 (52.3%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.3$)	62.58	84.14	47.54	55.41	144,429 (65.6%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.2$)	64.37	84.19	49.76	57.98	164,430 (74.7%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.1$)	66.05	84.24	53.04	60.09	186,353 (84.7%)
Ours (BLIP-2 Flan T5)	67.22	84.84	56.88	60.83	219,928 (100%)
Mini-Batch Size is 1					
UpDn† [2]	41.52	43.03	13.53	48.41	-
+ $\mathbf{I}_r$ ($\alpha = 0.1$)	49.84	43.61	28.37	59.01	107,569 (48.9%)
+ $\mathbf{I}_b$	51.26	76.08	13.95	48.49	51,497 (23.4%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.5$)	51.72	76.24	14.15	49.18	64,950 (29.5%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.3$)	54.03	76.44	16.74	52.52	99,751 (45.3%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.2$)	56.44	76.50	20.34	55.83	125,195 (56.9%)
+ $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.1$)	59.48	76.58	28.72	58.96	155,761 (70.8%)
Ours (BLIP-2 Flan T5)	67.22	84.84	56.88	60.83	219,928 (100%)

The MLLMs are BLIP-2 Flan T5. "# samples" denotes the number of samples MLLMs will forward.

distribution and only 17,163in the "Tail" distribution. 3) Incorporating both $\mathbf{I}_r$ and $\mathbf{I}_b$ leads to further performance gains, e.g., on the testdev set, $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.1$) outperforms the individual applications of $\mathbf{I}_b$ and $\mathbf{I}_r$ ($\alpha = 0.1$). These findings highlight the complementary nature of the two indicators. 4) Note that on the validation set, our two settings of collaborative inference (i.e., $\mathbf{I}_r$ ($\alpha = 0.1$) and $\mathbf{I}_b$ + $\mathbf{I}_r$ ($\alpha = 0.1$)) achieve better performance than forwarding with only MLLMs, while also significantly optimising computational resources (i.e., collaborative inference requiring only MLLMs forwards 49.3% and 78% of samples, respectively). 5) When encountering an extreme environment, i.e., only one sample can be accessed at each mini-batch, our findings are similar to those of mini-batches with sizes greater than 1. These results demonstrate the generality of our VQA-CI.

From the results in Table X regarding the VQA-CP v2 dataset, we have similar observations with those on the GQA-OOD dataset, except that $\mathbf{I}_b$ achieves much higher performance than $\mathbf{I}_r$ ($\alpha = 0.1$). This discrepancy underscores the predominance of bias issues over unreliability concerns. Moreover, the integration of these two indicators significantly enhances performance, as evidenced by $\mathbf{I}_b + \mathbf{I}_r$ ($\alpha = 0.1$) surpassing $\mathbf{I}_b$ by approximately 7%. The improvement demonstrates the importance and necessity of recognising biased and unreliable samples from the predictions of VQA specialists. In the extreme environment (i.e., batch size is 1), we experiment with the substitution of Gaussian noise for randomly selected samples to identify negative samples. However, this approach yielded inferior performance compared with the selection of random samples.

These findings suggest that employing Gaussian noise to identify biased samples is sub-optimal. Nevertheless, the application of both indicators leads to further performance enhancements.

In total, these results demonstrate the effectiveness of our collaborative inference strategy that integrates MLLMs and VQA specialists.

VI. Conclusion

In this paper, we have proposed a novel VQA framework that integrates MLLMs and VQA specialists for collaborative inference. Specifically, we devise biased and unreliable sample indicators to ascertain whether predictions from VQA specialists should be accepted. If a sample is deemed unbiased and reliable, its prediction is accepted. Otherwise, the sample is subject to re-evaluation using MLLMs. Moreover, we empirically find that VQA specialists contain rich knowledge that can aid the reasoning processes of MLLMs. To capitalise on this, we propose mining discriminative information (i.e., question-related image regions and candidate answers) from VQA specialists. The information is then seamlessly integrated into the prompts used for the reasoning processes of MLLMs. Extensive experiments on the GQA-OOD and VQA-CP v2 datasets demonstrate the effectiveness of our VQA-CI.

References

[1] R. Cadène, H. Ben-younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1989–1998.

[2] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[3] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.* 2015, pp. 2425–2433.

[4] S. Zhang et al., "LOIS: Looking out of instance semantics for visual question answering," *IEEE Trans. Multimedia.*, vol. 26, pp. 6202–6214, 2024.

[5] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter-and intra-modality interactions," *IEEE Trans. Multimedia*, vol. 23, pp. 3518–3529, 2020.

[6] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2020, pp. 7760–7768.

[7] Y. Xi et al., "A dynamic feature interaction framework for multi-task visual perception," *Int. J. Comput. Vis.*, vol. 131, no. 11, pp. 2977–2993, 2023.

[8] C. Cao, H. Zhang, Y. Lu, P. Wang, and Y. Zhang, "Scene-dependent prediction in latent space for video anomaly detection and anticipation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 224–239, 2025.

[9] A. Ebrahimi et al., "AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages," in *Proc. Annu. Meeting Assoc. Comput. Linguistics* 2022, pp. 6279–6299.

[10] DeepSeek-AI et al., "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," 2024, *arXiv:2405.04434*.

[11] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[12] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6325–6334.

[13] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1983–1991.

[14] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. Adv. Neural Inf. Process. Syst.,* 2018, pp. 1548–1558.

[15] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4971–4980.

[16] Z. Wen, G. Xu, M. Tan, Q. Wu, and Q. Wu, "Debiased visual question answering from feature and sample perspectives," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3784–3796.

[17] X. Zhu et al., "Overcoming language priors with self-supervised learning for visual question answering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 1083–1089.

[18] Y. Niu et al., "Counterfactual VQA: A cause-effect look at language bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12700–12710.

[19] Z. Wen et al., "Test-time model adaptation for visual question answering with debiased self-supervisions," *IEEE Trans. Multimedia*, vol. 26, pp. 2137–2147, 2024.

[20] H. W. Chung et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–70, 2024.

[21] L. Zheng et al., "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2023, vol. 36, pp. 46595–46623.

[22] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. in Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[23] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[24] Z. Yang et al., "An empirical study of GPT-3 for few-shot knowledge-based VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3081–3089.

[25] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.

[26] Z. Peng et al., "Grounding multimodal large language models to the world," in *Proc. Int. Conf. Learn. Representations*, 2024.

[27] W. Dai et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 49250–49267.

[28] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *Int. Conf. Learn. Representations*, 2024.

[29] R. Awal, L. Zhang, and A. Agrawal, "Investigating prompting techniques for zero-and few-shot visual question answering," 2023, *arXiv:2306.09996*.

[30] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf, "Roses are red, violets are blue... but should VQA expect them to?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2776–278.

[31] Y. Hu et al., "PromptCap: Prompt-guided image captioning for VQA with GPT-3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 2963–2975.

[32] R. Cadène, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "RUBi:Reducing unimodal biases for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 839–850.

[33] J. Yu et al., "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3196–3209, 2020.

[34] H. Zhong et al., "Self-adaptive neural module transformer for visual question answering," *IEEE Trans. Multimedia*, vol. 23, pp. 1264–1273, 2020.

[35] R. R. Selvaraju et al., "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *Proc. IEEE Int. Conf. Comput. Vis.* 2019, pp. 2591–2600.

[36] A. Vosoughi et al., "Cross modality bias in visual question answering: A causal view with possible worlds VQA," *IEEE Trans. Multimedia*, vol. 26, pp. 8609–8624, 2024.

[37] N. Ouyang et al., "Suppressing biased samples for robust VQA," *IEEE Trans. Multimedia*, vol. 24, pp. 3405–3415, 2022.

[38] Y. Song, X. Yang, Y. Wang, and C. Xu, "Recovering generalization via pre-training-like knowledge distillation for out-of-distribution visual question answering," *IEEE Trans. Multimedia*, vol. 26, pp. 837–851, 2024.

[39] Y. Pan, J. Liu, L. Jin, and Z. Li, "Unbiased visual question answering by leveraging instrumental variable," *IEEE Trans. Multimedia*, vol. 26, pp. 6648–6662, 2024.

[40] J. Wu and R. J. Mooney, "Self-critical reasoning for robust visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8601–8611.

[41] Y. Niu and H. Zhang, "Introspective distillation for robust question answering," in *Proc. Adv. Neural Inf. Process. Syst.* 2021, pp. 16292–16304.

[42] T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "MUTANT: A training paradigm for out-of-distribution generalization in visual question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 878–892.

[43] L. Chen et al., "Counterfactual samples synthesizing for robust visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10797–10806.

[44] L. Chen, Y. Zheng, and J. Xiao, "Rethinking data augmentation for robust visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 95–112.

[45] J. Kil, C. Zhang, D. Xuan, and W. Chao, "Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6346–6361.

[46] Z. Wen, Y. Wang, M. Tan, Q. Wu, and Q. Wu, "Digging out discrimination information from generated samples for robust visual question answering," in *Proc. Findings Assoc. Comput. Linguistics: ACL*, 2023, pp. 6910–6928.

[47] Z. Gao, X. Ao, X.-Y. Zhang, and C.-L. Liu, "Adapting vision-language models to open classes via test-time prompt tuning," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2024, pp. 439–452.

[48] C. Zhang, S. Stepputtis, K. P. Sycara, and Y. Xie, "Dual prototype evolving for test-time generalization of vision-language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 32111–32136.

[49] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[50] S. Luo et al., "Enhancing test time adaptation with few-shot guidance," 2024, *arXiv:2409.01341*.

[51] S. Niu et al., "Efficient test-time model adaptation without forgetting," in *Proc. Int. Conf. Mach. Learn.*, 2022, vol. 162, pp. 16888–16905.

[52] M. Boudiaf, R. Müller, I. B. Ayed, and L. Bertinetto, "Parameter-free online test-time adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8334–8343.

[53] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Proc. Int. Conf. Learn. Representations*, 2021.

[54] Y. Lan et al., "Improving zero-shot visual question answering via large language models with reasoning question prompts," in *Proc. ACM Int. Conf. Multimedia* 2023, pp. 4389–4400.

[55] X. Fu et al., "Generate Then Select: Open-Ended Visual Question Answering Guided by World Knowledge," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 2333–2346.

[56] Z. Wang, C. Chen, P. Li, and Y. Liu, "Filling the image information gap for VQA: Prompting large language models to proactively ask questions," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2023, pp. 2874–2890.

[57] J. Guo et al., "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10867–10877.

[58] K. Yang et al., "Good questions help zero-shot image reasoning," 2023, *arXiv:2312.01598*.

[59] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14974–14983.

[60] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 34892–34916.

[61] Q. Ye et al., "Mplug-Owl: Modularization Empowers Large Language Models With Multimodality," 2023, *arXiv:2304.14178*.

[62] OpenAI, "GPT-4 Technical Report," 2023, *arXiv:2303.08774*.

[63] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1571–1581.

[64] H. Ben-younes, R. Cadène, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8102–8109.

[65] S. Niu et al., "Towards stable test-time adaptation in dynamic wild world," in *Proc. Int. Conf. Learn. Representations*, 2023.

[66] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 4067–4080.

[67] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6700–6709.

[68] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[69] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 21–29.

[70] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 5099–5110.

Zhiquan Wen received the Ph.D. degree in software engineering from the South China University of Technology, Guangzhou, China, in 2024. He is currently a Postdoctoral Researcher with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China. His research interests include vision-and-language and test-time adaptation.

Mingkui Tan (Member, IEEE) received the Bachelor's Degree in environmental science and engineering in 2006 and master's degree in control science and engineering in 2009, both from Hunan University, Changsha, China, and the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014. He is currently a Professor with the School of Software Engineering, with South China University of Technology, Guangzhou, China. From 2014 to 2016, he worked as a Senior Research Associate on computer vision in the School of Computer Science, University of Adelaide, Australia. His research interests include machine learning, sparse analysis, deep learning and large-scale optimization.

Yaowei Wang (Member, IEEE) is currently a Professor with the Peng Cheng Laboratory and Harbin Institute of Technology, Shenzhen. He enjoys special government allowances of the State Council. He is the author or coauthor of more than 140 technical articles in international journals and conferences, including TOMM, ACM MM, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, and IJCAI. His current research interests include multimedia content analysis and understanding, machine learning, and computer vision. He was the Chair of the IEEE Digital Retina Systems Working Group and a Member of IEEE, CIE, CCF, CSIG. He was the recipient of the second prize of the National Technology Invention in 2017, first prize of the CIE Technology Invention in 2015, and first prize of the CIE Scientific and Technological Progress in 2022.

Qingyao Wu received the B.S. degree in software engineering from the South China University of Technology, Guangzhou, China, in 2007, and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2009 and 2013, respectively. He is currently a Professor with the School of Software Engineering, South China University of Technology. His current research interests include computer vision and data mining.

Qi Wu received the M.Sc. degree in Computer Science from the University of Bath, U.K., in 2011 and the Ph.D. degree in 2015. His educational background is primarily in computer science and mathematics. He is currently a Senior Lecturer (Assistant Professor) with the University of Adelaide, Australia. He is also an Associate Investigator with the Australia Centre for Robotic Vision. He is an ARC Discovery Early Career Researcher Award (DECRA) Fellow from 2019 to 2021. He works on vision and language problems, including image captioning, visual question answering, and visual dialog. His work has been authored or coauthored in prestigious journals and conferences such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, ICCV, AAAI, and ECCV.