Grounded Multi-modal Conversation for Zero-shot Visual Question Answering

Mohammad Reza Zarei[1*], Abbas Akkasi [1] and Majid Komeili[1]

Abstract— **Zero-shot visual question answering (VQA) poses a formidable challenge at the intersection of computer vision and natural language processing. Traditionally, this problem has been tackled using end-to-end pre-trained vision-language models (VLMs). However, recent advancements in large language models (LLMs) demonstrate their exceptional reasoning and comprehension abilities, making them valuable assets in multi-modal tasks, including zero-shot VQA. LLMs have been previously integrated with VLMs to solve zero-shot VQA in a conversation-based approach. However, while the focus in VQA tasks is often on specific regions rather than the entire image, this aspect has been overlooked in previous approaches. Consequently, the overall performance of the framework relies on the ability of the pre-trained VLM to locate the region of interest that is relevant to the requested visual information within the entire image. To address this challenge, this paper proposes Grounded Multi-modal Conversation for Zero-shot Visual Question Answering (GMC-VQA), a region-based framework that leverages the complementary strengths of LLMs and VLMs in a conversation-based approach. We employ a grounding mechanism to refine visual focus according to the semantics of the question and foster collaborative interaction between VLM and LLM, effectively bridging the gap between visual and textual modalities and enhancing comprehension and response generation for visual queries. We evaluate GMC-VQA across three diverse VQA datasets, achieving substantial average improvements of 10.04% over end-to-end VLMs and 2.52% over the state-of-the-art VLM-LLM communication-based framework, respectively. Our code is publicly available at `https://github.com/mrzarei5/GMC-VQA`.**

I. Introduction

Visual Question Answering (VQA) refers to a challenging task that lies at the intersection of language processing and image understanding and involves generating accurate textual answers to the questions about the visual content of an in image [1], [2]. This cognitive challenge demands a deep understanding of object interactions, relationships, actions, events, quantities, and textual elements present in the visual scene. Compared to other computer vision tasks like image captioning and text-to-image generation, VQA exhibits a greater level of complexity. This complexity stems from a variety of factors, including the diverse ways questions can be formulated, the vast array of relevant visual information, and the different types of questions that need to be answered.

Recent advancements in vision-language models (VLMs) have significantly impacted VQA tasks, including zero-shot VQA. However, VLMs face several challenges: 1) **Limited factual knowledge**: VLMs are primarily trained on image-text pairs, which might not contain sufficient factual knowledge to accurately answer complex questions in zero-shot scenarios. 2) **Limited understanding**: VLMs might not fully grasp common sense reasoning in zero-shot scenarios, leading to incorrect answers for questions that necessitate an understanding of everyday phenomena or established physical laws. 3) **Overreliance on visual cues**: In the context of zero-shot VQA, VLMs may overly rely on visual information, potentially overlooking logical reasoning based on textual context. 4) **Question/Image ambiguity**: VLMs can be confused by ambiguous questions, leading to incorrect or irrelevant answers. Furthermore, VLMs might struggle with images containing multiple objects or complex scenes, leading to difficulty in identifying the correct visual information.

While VLMs address these challenges by significantly increasing model sizes, utilizing extensive datasets, and employing advanced pre-training techniques [3]–[5], their size requires substantial infrastructure, making on-premise deployment difficult. For instance, Pixtral Large [4] demands over 300 GB of GPU RAM. Providers such as OpenAI, Google, and MistralAI offer serverless access through APIs; however, using APIs for multi-modal models in vision tasks can be expensive and raise privacy concerns due to the sensitive nature of image data, unlike text. As a result, leveraging VLMs with the possibility of local deployment is crucial for both managing costs and maintaining privacy in zero-shot VQA. However, improving the performance of such VLMs remains essential to fully address the complexities of VQA tasks, particularly in bridging gaps in reasoning and knowledge integration.

Recent studies have explored the possibility of leveraging large language models (LLMs) in VQA tasks [6]–[8] to overcome VQA challenges. LLMs have demonstrated exceptional capabilities in reasoning [9], [10]. Their proficiency in processing and comprehending intricate language structures enables them to effectively interpret question nuances, extract critical information, and construct logical reasoning chains. Moreover, LLMs excel at knowledge integration, allowing them to incorporate external and world knowledge into the VQA process [11], [12]. This capacity to reason over language and knowledge is indispensable for addressing questions that require inferential or deductive reasoning, significantly enhancing VQA system performance.

A key challenge when utilizing LLMs for VQA is effectively bridging the gap between textual and visual information, which is essential for enabling LLMs to accurately

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, and Digital Research Alliance of Canada.

[1] School of Computer Science, Carleton University, Ottawa, Canada

*Corresponding author: Mohammad Reza Zarei. E-mail: mohammadreza-zarei@cmail.carleton.ca.

interpret and understand image content [8]. The predominant solution to this challenge involves transforming images into textual representations, with image captioning serving as the most straightforward method for achieving this goal. However, image captions often lack the specific details necessary to accurately answer questions related to the image [13], [14]. To address this shortcoming, recent approaches have integrated the question into the image captioning process [14]. Nonetheless, such methods may still fail to capture all relevant details [8].

A promising advancement involves leveraging the LLM itself to inquire about essential visual information while utilizing a pre-trained VLM to provide the LLM with the requested visual details, thus preparing all necessary information for answering the initial question [7]. Although this LLM-VLM conversational framework has demonstrated effectiveness in zero-shot VQA scenarios, the information presented by VLM has been extracted from the entire image while both the initial VQA query and the subsequent requests from LLM typically focus on specific regions, rather than the entirety of the image. Therefore, the overall performance hinges on the VLM's capability to accurately identify regions of interest relevant to the requested visual information within the entire image.

This paper introduces Grounded Multi-modal Conversation for Zero-shot Visual Question Answering (GMC-VQA), a novel framework for zero-shot VQA. By incorporating image grounding, GMC-VQA enables the VLM to concentrate on image regions that are relevant to the given question. The model establishes an iterative question-answering conversation between the LLM and VLM for each identified region, facilitating the collection of region-specific visual information that aids LLM in generating a precise final answer to the initial query. Experimental results across three VQA datasets demonstrate the effectiveness of our proposed model, showing significant improvements compared to both end-to-end VLMs and previous VLM-LLM communication-based framework.

Our contributions are summarized as follows:

- We propose GMC-VQA, a multi-modal conversation-based approach which integrates a grounding mechanism to focus on specific regions of interest within images.
- GMC-VQA establishes an individual question-answering conversation between the pre-trained VLM and LLM for each relevant image region, resulting in enhanced targeted comprehension and response generation for visual queries.
- GMC-VQA outperforms existing end-to-end and conversation-based VQA approaches across three diverse VQA datasets.

II. Related work

A. Vision-Language Models

The advancement of vision-language pre-training models has led to significant achievements in vision-language tasks. VLMs aim to learn alignments between visual and textual information using large-scale image-text pairs, then use the model in zero-shot setting or fine-tune it on downstream tasks [15]–[17]. Based on their capabilities and initial objectives, VLMs can be categorized as either discriminative or generative models. Discriminative VLMs [18], [19] focus on learning robust representations for both image and text, facilitating recognition and matching tasks. In contrast, generative VLMs [20], [21] concentrate on tasks that require text or image generation capabilities, such as image captioning and visual question answering. Recent VLMs with text generation capabilities support multiple text generation tasks. Among these models, LLaVA [20] trained an end-to-end large multimodal model that connects a vision encoder with an LLM for general-purpose visual and language understanding. BLIP-2 [22] bridged the modality gap between image and text with a lightweight Querying Transformer, pre-trained in two stages, bootstrapping vision-language representation learning from a frozen image encoder and bootstrapping vision-to-language generative learning from a frozen language model in the first and the second stages, respectively. In GMC-VQA, we leverage the capability of pre-trained VLMs in image captioning and question-answering, suitable for local deployment.

B. Visual Question Answering

A variety of methods have been proposed for VQA in recent years. Li et al. [23] explored three VQA approaches: GAN-based methods generated answer embeddings but struggled with complexity; autoencoders learned question and image embeddings, performing comparably; and attention mechanisms with MCB addressed language priors and attention modeling, balancing complexity and performance. Bao et al. [24] approached the VQA problem with a confidence-based neural-symbolic model which evaluates the uncertainty of neural network inferences and uses that information to guide reasoning. Yu et al. [25] proposed a graph-based recurrent reasoning network (GRUC) for VQA, leveraging cross-modal knowledge reasoning through graph-structured multimodal knowledge representations. ViLT, presented in [26], is a minimal model that incorporates text embeddings into a Vision Transformer. This model converts VQA task to a classification task. While the mentioned models focus on designing and training a more powerful VQA model, we propose a training-free VQA framework aimed at enhancing zero-shot performance by leveraging the reasoning and understanding capabilities of an external LLM.

C. Prompting Large Language Models for Visual Question Answering

Reasoning and understanding capabilities of LLMs have recently been used in VQA. Liang et al. [27] introduced a novel approach for knowledge-based VQA that leverages LLMs to actively and progressively gather visual information in a task-oriented manner. This approach involves generating initial hypotheses, collecting relevant visual evidence, and

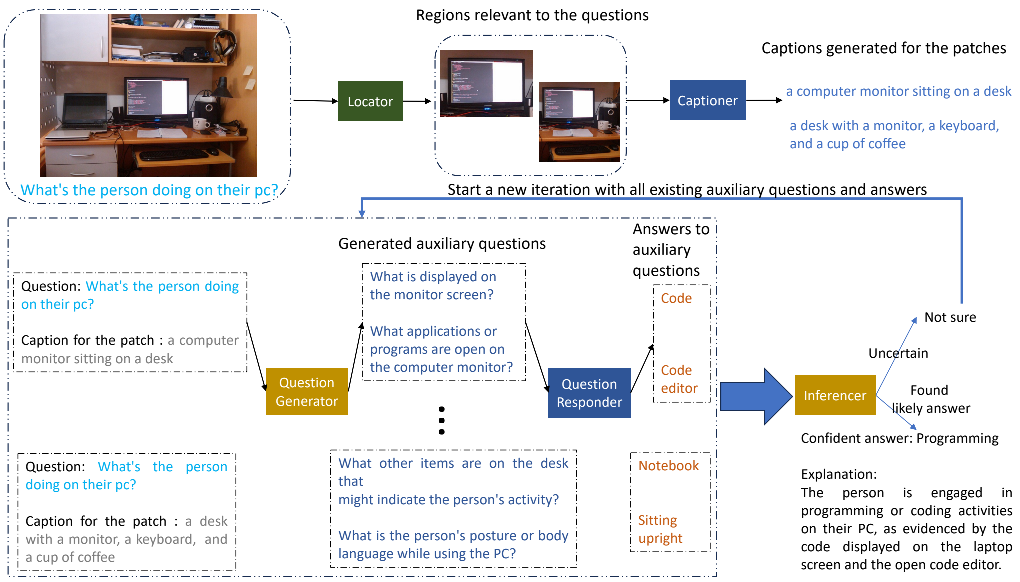

What's the person doing on their pc?
Locator
Regions relevant to the questions
Captioner
Captions generated for the patches
a computer monitor sitting on a desk
a desk with a monitor, a keyboard, and a cup of coffee
Start a new iteration with all existing auxiliary questions and answers
Generated auxiliary questions
Answers to auxiliary questions
Question: What's the person doing on their pc?
Caption for the patch : a computer monitor sitting on a desk
Question Generator
What is displayed on the monitor screen?
What applications or programs are open on the computer monitor?
Question Responder
Code
Code editor
Inferencer
Not sure
Uncertain
Found likely answer
Confident answer: Programming
Explanation:
The person is engaged in programming or coding activities on their PC, as evidenced by the code displayed on the laptop screen and the open code editor.
Question: What's the person doing on their pc?
Caption for the patch : a desk with a monitor, a keyboard, and a cup of coffee
What other items are on the desk that might indicate the person's activity?
What is the person's posture or body language while using the PC?
Notebook
Sitting upright

Fig. 1. Overview of the GMC-VQA framework for zero-shot VQA. The Locator extracts relevant image patches, which are then captioned by the Captioner. An iterative dialogue between the Question Generator and Question Responder generates auxiliary questions and answers for each patch to aid in resolving the initial query. After each iteration, the Inferencer synthesizes the information and attempts to infer the answer with high confidence; if unsuccessful, the process continues with additional iterations.

verifying these hypotheses through multiple rounds of reasoning. Lan et al. [6] presented a method to enhance zero-shot VQA by leveraging reasoning question prompts. The approach involves generating self-contained questions that clarify the original ambiguous queries, enabling LLMs to better understand and answer them. Img2LLM [28] enables frozen LLMs to perform zero-shot VQA by first extracting candidate answers from image captions and then generating corresponding synthetic questions, along with question-relevant captions, to construct effective prompts. LAMOC [29] improves zero-shot VQA by training a captioning model to generate task-aware image descriptions using feedback from a frozen language model, enhancing its relevance to the question and utility for answer prediction. Wang et al. [8] enables LLM to proactively ask questions and gather more detailed information about images. A refinement module is then adopted to summarize the collected information, allowing the LLM to predict the answer based on the summarized data. IdealGPT [7] also leverages LLM-VLM communication, introducing an iterative framework in which an LLM asks supporting questions in multiple iterations, and a pre-trained VLM responds to those questions concerning the image. GMC-VQA also lies in communication-based VLM-LLM frameworks. However, unlike the previous methods, we enable our framework to concentrate on image regions that are relevant to the given question by leveraging image grounding and establishing an individual LLM-VLM conversation for each region.

III. Method

A. Overview

In this section, we introduce GMC-VQA, a novel framework for zero-shot VQA. The model is comprised of four key modules: Locator, Question Generator, Question Responder and Inferencer. Given the query image I and the question q, a visual grounding model (Locator, Section III-B) is leveraged to identify relevant patches in I pertaining to q. Then, an iterative question-answering conversation is initiated between a pre-trained LLM (Question Generator, Section III-C) and a pre-trained VLM (Question Responder, III-D) to extract grounded knowledge relevant to the q for each patch. Finally, a pre-trained LLM (Inferencer, Section III-E) synthesizes the information from all patches to infer the answer to the question q. The iterative conversation between the Question Generator and Question Responder continues until the Inferencer can infer the answer to the initial question with high confidence or until a predefined limit on the number of iterations is reached. The framework is presented in Fig. 1.

B. Locator

In vqa tasks, the question q pertaining to the query image may often focus on specific regions rather than the entire image. This specificity can pose challenges for the VQA model in accurately associating the question with the relevant parts of the image and providing an accurate answer. To address this challenge, we employ a pre-trained open-set visual grounding model as the Locator module. This module receives q and I as the input, associates q with key object regions of the image, and extracts the relevant patches. The patches $\mathcal{P} = [p_1; p_2; ...; p_k]$ are identified by the grounding model L as:

$$\mathcal{P} = L(I, q) \tag{1}$$

The number of extracted regions is dynamic, varying from one sample to another. We also treat the original image as an individual patch and append it to $\mathcal{P}$.

C. Question Generator

LLMs have shown strong capability in following instructions and performing reasoning tasks [9], [10]. We harness this capability to generate questions about each patch that complements the original question q. These auxiliary questions are collected in multiple iterations through a communication with Question Responder and provide evidence for the Inferencer module to answer the initial question q. For each patch $p_i \in \mathcal{P}$, we first utilize an image captioning model M to generate a caption for p_i:

$$c_i = M(p_i) \tag{2}$$

This caption serves as the preliminary information for the patch presented to a pre-trained LLM, acting as the Question Generator G, along with the original question q to generate auxiliary questions that help gather supplementary information for answering the original question. We also present the auxiliary questions generated in the previous interactions and their corresponding answers to the Question Generator, to prevent it from generating repetitive questions and aid in generating questions that complement the previous ones. The questions for the patch p_i in iteration t are generated as:

$$\mathcal{Q}_i{}^{(t)} = G(q, c_i, \mathcal{Q}_i, \mathcal{A}_i) \tag{3}$$

where $\mathcal{Q}_i$ and $\mathcal{A}_i$ are the set of questions and their corresponding answers for the patch p_i from all previous iterations. These sets are empty in the first iteration. Since some patches may be less relevant to the initial question q, we ask LLM to limit the number of generated questions in this case.

At the end of the iteration, $\mathcal{Q}_i$ and $\mathcal{A}_i$ are updated by appending $\mathcal{Q}_i{}^{(t)}$ and $\mathcal{A}_i{}^{(t)}$ to them, respectively:

$$\mathcal{Q}_i \leftarrow \mathcal{Q}_i \cup \mathcal{Q}_i{}^{(t)}, \mathcal{A}_i \leftarrow \mathcal{A}_i \cup \mathcal{A}_i{}^{(t)} \tag{4}$$

Where $\mathcal{A}_i{}^{(t)}$ is the set of answers to the questions generated in the iteration t. The answering process is presented in the next section.

D. Question Responder

The Question Responder answers the auxiliary questions $\mathcal{Q}_i{}^{(t)}$ generated by the Question Generator for the patch p_i in iteration t. We use a pre-trained VLM as the Question Responder without any further fine-tuning. Each auxiliary question $q' \in \mathcal{Q}_i{}^{(t)}$ is answered individually by the Question Responder:

$$a_{q'} = R(q', p_i) \tag{5}$$

E. Inferencer

We utilize the reasoning capabilities of pre-trained LLMs to synthesize information from all patches and infer the answer to the initial question q. At the end of each iteration, the initial question q is presented alongside a collection of information from each patch, formatted as $[c_i, \mathcal{Q}_i, \mathcal{A}_i]$ for $i = 1, 2, \ldots, k+1$. Here, c_i represents the caption for the i-th patch, $\mathcal{Q}_i$ denotes the generated auxiliary questions related to that patch, and $\mathcal{A}_i$ contains their corresponding answers. This aggregated information is then provided to a pre-trained LLM designated as Inferencer F. The module is tasked with delivering the answer to the initial question a if it is confident in its response, as well as detailing the reasoning and inference process $r_{explain}$ based on the supplied information:

$$a, r_{explain} = F\left(q, \{[c_i, \mathcal{Q}_i, \mathcal{A}_i]\}_{i=1}^{k+1}\right) \tag{6}$$

If Inferencer is uncertain about the answer and the predefined iteration limit is not reached, the procedure continues with a new iteration. The framework pipeline is summarized in Algorithm 1.

Algorithm 1 Pipeline of the proposed method

```
Input: Query image $I$, the question $q$ and the maximum iteration number $max\_iter$
Output: Answer to $q$
  $\mathcal{P} = L(I, q)$ {Locate the patches $\mathcal{P} = [p_1; p_2; ...; p_k]$ in image $I$ relevant to the query $q$}
  $\mathcal{P} \leftarrow \mathcal{P} \cup I$ {Append the initial image to set of patches}
  $\mathcal{C} = [\ ]$, $\mathcal{Q} = [\ ]$, $\mathcal{A} = [\ ]$ {Initialize captions, questions and answers list}
  for all $p_i$ in $\mathcal{P}$ do
    $c_i = M(p_i)$ {Generate caption}
    $\mathcal{C} \leftarrow \mathcal{C} \cup c_i$
    $\mathcal{Q}_i = [\ ]$, $\mathcal{A}_i = [\ ]$ {initialize patch-specific questions and answers sets}
  end for
  $t = 0$ {Iteration number}
  repeat
    $t = t + 1$
    for all $p_i$ in $\mathcal{P}$ do
      $\mathcal{Q}_i{}^{(t)} = G(q, c_i, \mathcal{Q}_i, \mathcal{A}_i)$ {Generate questions}
      $\mathcal{A}_i{}^{(t)} = [\ ]$
      for all $q'$ in $\mathcal{Q}_i{}^{(t)}$ do
        $a_{q'} = R(q', p_i)$ {Answer generated questions}
        $\mathcal{A}_i{}^{t} \leftarrow \mathcal{A}_i{}^{t} \cup a_{q'}$
      end for
      $\mathcal{Q}_i \leftarrow \mathcal{Q}_i \cup \mathcal{Q}_i{}^{(t)}, \mathcal{A}_i \leftarrow \mathcal{A}_i \cup \mathcal{A}_i{}^{(t)}$
    end for
    $a, r_{explain} = F\left(q, \{[c_i, \mathcal{Q}_i, \mathcal{A}_i]\}_{i=1}^{k+1}\right)$ {Infer the answer to $q$ and provide the inference process}
  until $t$ == $max\_iter$ or $a$ != 'not sure'
```

IV. Experiments and Evaluation

A. Experimental Setup

Datasets and experimental settings. We conduct experiments on three VQA datasets, VQAv2 [30], OK-VQA [31], and ST-VQA [32], covering diverse VQA challenges. VQAv2 focuses on vision, language, and commonsense reasoning using COCO images, while OK-VQA extends this by requiring external knowledge beyond the visuals. ST-VQA focuses on understanding scene text and comprises images from various sources like COCO-Text [33] and ICDAR2015 [34]. For each dataset, 2,000 samples were randomly selected for zero-shot evaluation: validation and test sets for VQAv2 and OK-VQA, respectively, and the training set for ST-VQA (which lacks labeled test data).

Implementation details. We employ Grounding DINO [35], an open-set object detector, as the Locator module. For the Question Responder, we utilize three practically deployable vision-language models (VLMs): BLIP-2 (FlanT5-XL), InstructBLIP (FlanT5-XL) (each with 3.4 billion parameters), and LLaVA (LLaMA-2 7B) (7.3 billion parameters). The size of the object detector is negligible compared to the VLMs, comprising only 172 million parameters. In each experiment, the same VLM used as the Question Responder also serves as the Image Captioning model. For the Question Generator and Inferencer, we primarily use GPT-4o mini due to its affordability and strong performance. We also evaluate our model with GPT-3.5-Turbo and three open-source LLMs: Llama 2 13B, Mistral 7B Instruct v0.2, and Mixtral 8x7B Instruct v0.1. In each experiment, the same LLM is used for both generating questions and inference. All models in our framework are pre-trained and used without further fine-tuning. The temperature of all LLMs is set to 0 to ensure reproducibility. Following IdealGPT [7], four questions are generated and asked per iteration, with the maximum number of iterations max_iter fixed at four for all experiments. The base prompts for GMC-VQA are adapted from [7] and modified to fit our framework. All experiments are conducted on a machine equipped with an NVIDIA RTX 3090 GPU.

Evaluation method. While exact string matching accuracy has long been the primary metric for automatically evaluating VQA models, this approach is no longer suitable due to the recent shift in VQA research towards zero-shot transfer and the increased diversity in the formats of generated answers [36]. To address this limitation and inspired by the recent success of LLMs as evaluators, referred to as LLM-as-a-Judge [37], [38], we evaluate the accuracy of all approaches using GPT-4o mini. Specifically, we assess whether a generated answer is semantically aligned with the ground truth answer. We also request intermediate reasoning steps from GPT 4o mini to further enhance evaluation reliability [39].

Comparable methods. In addition to comparing GMC-VQA with off-the-shelf VLMs used for VQA, including BLIP-2 (FlanT5-XL), InstructBLIP (FlanT5-XL), and LLaVA (LLaMA-2 7B), we also evaluate it against IdealGPT [7], which leverages VLM-LLM communication for zero-shot VQA. The VLMs were chosen for their moderate size that enables deployment on local machine. Additionally, these VLMs are the same as ones used within our framework as the Question Responder, ensuring fairness in our comparisons. Furthermore, we maintain the same LLM/VLM configurations for both our model and IdealGPT across experiments to further ensure a fair comparison

B. VQA Performance Comparison

We compare the zero-shot VQA accuracy of GMC-VQA with baselines on VQAv2, OK-VQA and ST-VQA in Table I. The results of our model and IdealGPT are reported with GPT-4o mini as the LLM, and the VLM with the best accuracy from LLaVA, BLIP-2 and InstructBLIP. VLM-specific results are presented in the next section.

TABLE I
ACCURACY COMPARISON BETWEEN GMC-VQA AND COMPETITORS ON THREE VQA DATASETS.

Method	VQAv2	OK-VQA	ST-VQA
LLaVA	50.5	44.65	24.1
BLIP-2	51.3	41.85	22.2
InstructBLIP	52.1	41.65	23.1
IdealGPT	59.7	50.25	30
GMC-VQA	**61.8**	**51.85**	**33.9**

As demonstrated, the proposed model consistently outperforms all baseline VLMs across the three datasets. Moreover, it surpasses the IdealGpt by an average margin of 2.52%. The performance gap between the LLM-VLM communication-based models and the VLM baselines underscores the value of integrating LLMs with VLMs. Moreover, the proposed model's superior performance relative to the IdealGPT highlights the efficacy of grounding in enhancing overall performance.

C. Performance Comparison with IdealGPT: Effect of VLM Variation

We evaluated GMC-VQA with different VLMs and compared it to IdealGPT, a communication-based framework, as shown in Table II. In all experiments, GPT-4o Mini was used as the LLM. GMC-VQA achieved the best accuracy on VQAv2 and OK-VQA when paired with InstructBLIP, likely due to its ability to handle complex questions and reason over image content. On ST-VQA, however, BLIP-2 performed better, as this dataset focuses heavily on text recognition and interpretation. In this case, the Locator identified relevant regions, while the VLM primarily handled text extraction.

Across all three datasets, GMC-VQA consistently outperformed IdealGPT, with accuracy improvements of 2.43%, 1.98%, and 6.25% on VQAv2, OK-VQA, and ST-VQA, respectively. The larger improvement on ST-VQA highlights the importance of precise region grounding and accurate text understanding, areas where IdealGPT struggles compared to GMC-VQA's region-based approach.

D. Performance Comparison with IdealGPT: Effect of LLM Variation

We evaluated GMC-VQA with various commercial and open-source LLMs, including GPT-4o Mini, ChatGPT, Mistral 7B Instruct v0.2, Mixtral 8x7B Instruct v0.1, and Llama 2 (13B), and compared its performance to IdealGPT (Table III). Given the considerably better results achieved with GPT-4o Mini and ChatGPT, we focused on reporting results for all LLMs with BLIP-2, while results for LLaVA and InstructBLIP are shown only with GPT-4o Mini and ChatGPT.

Across all tested LLMs, GMC-VQA consistently outperformed IdealGPT, which can be attributed to its region-based LLM-VLM communication. The best overall performance was achieved using GPT-4o Mini, demonstrating its strong capabilities despite being a cost-effective option. While commercial LLMs like GPT-4o Mini and ChatGPT delivered

TABLE II

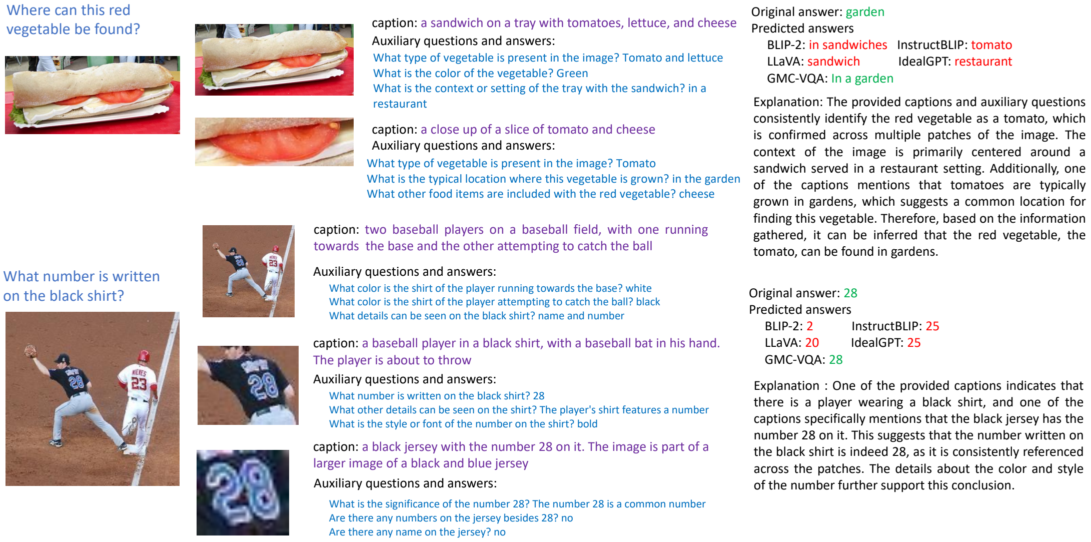
PERFORMANCE COMPARISON BETWEEN GMC-VQA AND IDEALGPT USING VARIOUS VLMS AND GPT-4O MINI AS LLM.

VLM	VQAv2		OK-VQA		ST-VQA	
	IdealGPT	GMC-VQA	IdealGPT	GMC-VQA	IdealGPT	GMC-VQA
LLaVA	54.8	**56.35**	49.6	**50.75**	30	**33.15**
BLIP-2	55.65	**59.3**	46.6	**49.8**	22.65	**33.9**
InstructBLIP	59.7	**61.8**	50.25	**51.85**	26.1	**30.45**

TABLE III

PERFORMANCE COMPARISON OF GMC-VQA AND IDEALGPT ACROSS VARIOUS LLMS. RESULTS FOR LLAVA AND INSTRUCTBLIP ARE SHOWN ONLY WITH CHATGPT AND GPT-4O MINI, AS THESE LLMS OUTPERFORMED OTHERS WHEN USED WITH BLIP-2.

VLM	LLM	VQAV2		OKVQA		STVQA	
		IdealGPT	Ours	IdealGPT	Ours	IdealGPT	Ours
BLIP-2	Mistral 7B Instruct v0.2	48.6	**49.4**	38	**39.95**	20.1	**27.4**
	Mixtral 8x7B Instruct v0.1	46.95	**48.95**	**43.35**	41.3	19.4	**28.7**
	Llama 2 13b	42	**47.4**	**38.85**	36.05	16.1	**21.2**
	ChatGPT	53.2	**55**	46.35	**47.8**	21.4	**30.05**
	GPT-4o mini	55.65	**59.3**	46.6	**49.8**	22.65	**33.9**
LLaVA	ChatGPT	50.95	**51.55**	49.5	**49.7**	27.05	**30.1**
	GPT-4o mini	54.8	**56.35**	49.6	**50.75**	30	**33.15**
InstructBLIP	ChatGPT	55.4	**57.1**	48.1	**48.8**	21.15	**29.6**
	GPT-4o mini	59.7	**61.8**	50.25	**51.85**	26.1	**30.45**



Where can this red vegetable be found?

caption: a sandwich on a tray with tomatoes, lettuce, and cheese
Auxiliary questions and answers:
What type of vegetable is present in the image? Tomato and lettuce
What is the color of the vegetable? Green
What is the context or setting of the tray with the sandwich? in a restaurant

caption: a close up of a slice of tomato and cheese
Auxiliary questions and answers:
What type of vegetable is present in the image? Tomato
What is the typical location where this vegetable is grown? in the garden
What other food items are included with the red vegetable? cheese

Original answer: garden
Predicted answers
BLIP-2: in sandwiches InstructBLIP: tomato
LLaVA: sandwich IdealGPT: restaurant
GMC-VQA: In a garden

Explanation: The provided captions and auxiliary questions consistently identify the red vegetable as a tomato, which is confirmed across multiple patches of the image. The context of the image is primarily centered around a sandwich served in a restaurant setting. Additionally, one of the captions mentions that tomatoes are typically grown in gardens, which suggests a common location for finding this vegetable. Therefore, based on the information gathered, it can be inferred that the red vegetable, the tomato, can be found in gardens.

What number is written on the black shirt?

caption: two baseball players on a baseball field, with one running towards the base and the other attempting to catch the ball
Auxiliary questions and answers:
What color is the shirt of the player running towards the base? white
What color is the shirt of the player attempting to catch the ball? black
What details can be seen on the black shirt? name and number

caption: a baseball player in a black shirt, with a baseball bat in his hand. The player is about to throw
Auxiliary questions and answers:
What number is written on the black shirt? 28
What other details can be seen on the shirt? The player's shirt features a number
What is the style or font of the number on the shirt? bold

caption: a black jersey with the number 28 on it. The image is part of a larger image of a black and blue jersey
Auxiliary questions and answers:
What is the significance of the number 28? The number 28 is a common number
Are there any numbers on the jersey besides 28? no
Are there any name on the jersey? no

Original answer: 28
Predicted answers
BLIP-2: 2 InstructBLIP: 25
LLaVA: 20 IdealGPT: 25
GMC-VQA: 28

Explanation : One of the provided captions indicates that there is a player wearing a black shirt, and one of the captions specifically mentions that the black jersey has the number 28 on it. This suggests that the number written on the black shirt is indeed 28, as it is consistently referenced across the patches. The details about the color and style of the number further support this conclusion.

Fig. 2. Qualitative examples of our GMC-VQA. GMC-VQA can effectively focus on important regions, extract relevant visual information from those regions through a region-based conversation, and infer the final answer from the information collected from different patches.

the highest gains, GMC-VQA exhibited the most significant improvements on VQAv2 and ST-VQA (2.18% and 6.73%, respectively) compared to OK-VQA (0.6%). This highlights the influence of dataset characteristics—OK-VQA requires external knowledge beyond the image and benefits less from region-based reasoning compared to the visually focused VQAv2 and text-heavy ST-VQA datasets.

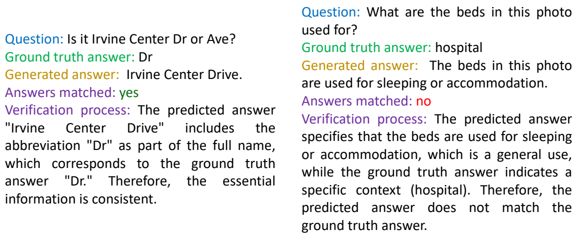
E. Qualitative Examples

To illustrate how GMC-VQA performs in zero-shot VQA, we present two examples in Fig. 2, with one from OK-VQA (top) and the other from ST-VQA (bottom). Both examples demonstrate how GMC-VQA effectively processes image regions in just one iteration of question-answering communication.

In the first example, the image shows a sandwich with tomato, lettuce, and cheese, and the question asks, ”Where can this red vegetable be found?” The Locator identifies a patch highlighting a tomato slice in the sandwich. During the conversation, while the full image provides a distracting context (”in a restaurant”), the patch-specific communication focuses on the tomato, inferring its growing location as ”in a garden.” Unlike other models that failed to answer correctly, GMC-VQA integrates region-specific insights, even leveraging complementary decisions from multiple patches when some information is inaccurate.

In the second example, two baseball players are pictured, one in a black shirt and the other in white, with the question asking for the number displayed on the black shirt. Competing models fail to locate the target region and identify the text simultaneously. In contrast, GMC-VQA successfully extracts the relevant patch of the black shirt, allowing the pre-trained VLM to provide the correct number during the region-based communication.

These examples demonstrate GMC-VQA's ability to focus on critical regions, extract relevant information, and combine



Question: Is it Irvine Center Dr or Ave?
Ground truth answer: Dr
Generated answer: Irvine Center Drive.
Answers matched: yes
Verification process: The predicted answer "Irvine Center Drive" includes the abbreviation "Dr" as part of the full name, which corresponds to the ground truth answer "Dr." Therefore, the essential information is consistent.

Question: What are the beds in this photo used for?
Ground truth answer: hospital
Generated answer: The beds in this photo are used for sleeping or accommodation.
Answers matched: no
Verification process: The predicted answer specifies that the beds are used for sleeping or accommodation, which is a general use, while the ground truth answer indicates a specific context (hospital). Therefore, the predicted answer does not match the ground truth answer.

Fig. 3. Examples of performance evaluation using GPT-4o Mini, illustrating its ability to accurately match VQA model answers with ground truth and provide a detailed verification process.

insights to produce precise, context-aware answers.

F. Performance Analysis

We analyzed GMC-VQA's performance to understand the kinds of questions it answers correctly compared to IdealGPT and to identify cases where errors persist. Unlike IdealGPT, which relies solely on the entire image for its question-answering dialogue, GMC-VQA primarily addresses errors related to questions that require attention to specific parts of the image. We found that when multiple objects are present, the VLM struggles to provide accurate answers to questions that ask for detailed information about a specific object or part of the image. As a result, IdealGPT often has difficulty generating precise responses.

However, there are specific cases where GMC-VQA fails to correct the errors present in IdealGPT. For instance, when the Locator is unable to extract any patches, our model defaults to a question-answering process based on the entire image, similar to IdealGPT. Consequently, the answer will depend on information from the whole image, leaving the error unresolved. This often happens when the original question does not refer to a specific object in the image, making it impossible for the grounding model to identify any relevant objects. For example, the question ”Is this in America?” does not point to a particular object, so the Locator cannot extract any relevant patches.

G. Validating LLM Performance in VQA Accuracy Assessment

We used GPT-4o Mini to assess model accuracy in zero-shot VQA task. Since this is the same LLM employed in our model, we conducted a manual validation to investigate potential evaluation bias. Specifically, we randomly selected 100 samples from each dataset, using InstructBLIP as the VLM and GPT-4o Mini as the LLM, and manually reviewed the evaluation results. Our analysis revealed that GPT-4o Mini demonstrates human-level performance as an evaluator for VQA: for VQAv2 and ST-VQA, there were no discrepancies between the model's and human judgments, while only one partial mismatch was found in OK-VQA's validation set. These findings confirm the reliability of GPT-4o Mini for automatic evaluation and highlight its strong reasoning and comprehension capabilities. Example evaluations are shown in Fig. 3.

TABLE IV
IMPACT OF VARYING NOISE LEVELS ON GMC-VQA ACCURACY, DEMONSTRATING THE ROLE OF BOUNDING BOX PRECISION IN THE LOCATOR MODULE'S PERFORMANCE.

Noise level	VQAv2	OK-VQA	ST-VQA
0	61.8	51.85	30.45
0.2	61.5	51.05	29.5
0.5	60.75	50.9	27.6
0.8	59.25	50.4	25.95

TABLE V
AVERAGE NUMBER OF ITERATIONS PER QUESTION USING GMC-VQA, WITH 95% CONFIDENCE INTERVALS. GPT-4O MINI IS USED AS LLM.

VLM	VQAV2	OKVQA	STVQA
LLaVA	1.21 ± 0.02	1.14 ± 0.02	1.37 ± 0.03
BLIP-2	1.36 ± 0.03	1.42 ± 0.04	1.81 ± 0.05
InstructBLIP	1.51 ± 0.04	1.56 ± 0.09	1.96 + 0.05

H. Locator Sensetivity Analysis

To examine GMC-VQA's sensitivity to patches identified by the Locator, we conducted experiments introducing controlled noise into the bounding box coordinates. Noise was generated as a randomized adjustment proportional to the bounding box dimensions, drawn from a uniform distribution over $(-1, 1)$ and scaled by a noise level parameter. The modified coordinates were clamped to remain within image boundaries, and degenerate boxes (e.g., collapsing into a line or point) were discarded.

Table IV summarizes the results. As noise levels increased, the model's accuracy declined slightly, underscoring the importance of precise grounding. However, the performance drop remained small in most cases, as the model can still utilize partially accurate patches. Additionally, GMC-VQA leverages the original image as a fallback source of information, enabling it to maintain a baseline level of performance even when patches are noisy or imprecise.

I. Experiment on Number of Iterations

We report the average number of iterations used to answer each question using GMC-VQA on each dataset with 95% confidence interval in Table V. While the maximum iteration number max_iter is set to 4 in all our experiments, we can understand from the averages and confidence intervals that it is more than adequate for the majority of samples.

V. CONCLUSION

In this paper, we introduced Grounded Multi-modal Conversation for Zero-shot Visual Question Answering (GMC-VQA), a framework that effectively combines the strengths of VLMs and LLMs by initializing region-based conversations between VLM and LLM. By incorporating a grounding mechanism, GMC-VQA enables the selective focus on relevant image regions based on the semantics of the questions, improving VQA performance. Experiments across three VQA datasets demonstrated performance enhancements over both traditional end-to-end VLMs and existing VLM-LLM communication framework.

References

[1] A. Nada and M. Chen, "Visual question answering," in *2024 International Conference on Computing, Networking and Communications (ICNC)*, 2024, pp. 6–10.

[2] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, "Visual question answering using deep learning: A survey and performance analysis," in *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*. Springer, 2021, pp. 75–86.

[3] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[4] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet *et al.*, "Pixtral 12b," *arXiv preprint arXiv:2410.07073*, 2024.

[5] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[6] Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian, "Improving zero-shot visual question answering via large language models with reasoning question prompts," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4389–4400.

[7] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. A. Ayyubi, K.-W. Chang, and S.-F. Chang, "Idealgpt: Iteratively decomposing vision and language reasoning via large language models," *arXiv preprint arXiv:2305.14985*, 2023.

[8] Z. Wang, C. Chen, P. Li, and Y. Liu, "Filling the image information gap for vqa: Prompting large language models to proactively ask questions," *arXiv preprint arXiv:2311.11598*, 2023.

[9] Z. Li, Y. Cao, X. Xu, J. Jiang, X. Liu, Y. S. Teo, S.-w. Lin, and Y. Liu, "Llms for relational reasoning: How far are we?" *arXiv preprint arXiv:2401.09042*, 2024.

[10] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

[11] T.-H. Wu, G. Biamby, J. Quenum, R. Gupta, J. E. Gonzalez, T. Darrell, and D. M. Chan, "Visual haystacks: Answering harder questions about sets of images," *arXiv preprint arXiv:2407.13766*, 2024.

[12] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," *arXiv preprint arXiv:2210.02928*, 2022.

[13] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of gpt-3 for few-shot knowledge-based vqa," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3081–3089.

[14] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided task-aware image captioning," *arXiv preprint arXiv:2211.09699*, 2022.

[15] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

[16] J. Wang, X. Hu, Z. Gan, Z. Yang, X. Dai, Z. Liu, Y. Lu, and L. Wang, "Ufo: A unified transformer for vision-language representation learning," *arXiv preprint arXiv:2111.10023*, 2021.

[17] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5579–5588.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[19] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[21] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[23] P. Li, Q. Yang, X. Geng, W. Zhou, Z. Ding, and Y. Nian, "Exploring diverse methods in visual question answering," *arXiv preprint arXiv:2404.13565*, 2024.

[24] Y. Bao, T. Xing, and X. Chen, "Confidence-based interactable neural-symbolic visual question answering," *Neurocomputing*, vol. 564, p. 126991, 2024.

[25] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, "Cross-modal knowledge reasoning for knowledge-based visual question answering," *Pattern Recognition*, vol. 108, p. 107563, 2020.

[26] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594.

[27] M. Liang, Y. Wu *et al.*, "Toa: task-oriented active vqa," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[28] J. Guo, J. Li, D. Li, A. M. Huat Tiong, B. Li, D. Tao, and S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 867–10 877.

[29] Y. Du, J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Zero-shot visual question answering with language model feedback," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9268–9281.

[30] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[31] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[32] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas, "Scene text visual question answering," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4291–4301.

[33] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," 2016. [Online]. Available: https://arxiv.org/abs/1601.07140

[34] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE, 2015, pp. 1156–1160.

[35] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[36] O. Mañas, B. Krojer, and A. Agrawal, "Improving automatic vqa evaluation using large language models," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[37] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[38] S. Lee, S. Kim, S. Park, G. Kim, and M. Seo, "Prometheus-vision: Vision-language model as a judge for fine-grained evaluation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 11 286–11 315.

[39] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.