electronics

MDPI

Article

LCV2: A Universal Pretraining-Free Framework for Grounded Visual Question Answering

Yuhan Chen [1], Lumei Su [1,2,*], Lihua Chen [1] and Zhiwei Lin [1]

[1] School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen 361024, China; 2222031340@stu.xmut.edu.cn (Y.C.); 2322131006@stu.xmut.edu.cn (L.C.); lzw7023@163.com (Z.L.)
[2] Xiamen Key Laboratory of Frontier Electric Power Equipment and Intelligent Control, Xiamen 361024, China
* Correspondence: sulumei@163.com

Abstract: Grounded Visual Question Answering systems place heavy reliance on substantial computational power and data resources in pretraining. In response to this challenge, this paper introduces the LCV2 modular approach, which utilizes a frozen large language model (LLM) to bridge the off-the-shelf generic visual question answering (VQA) module with a generic visual grounding (VG) module. It leverages the generalizable knowledge of these expert models, avoiding the need for any large-scale pretraining. Innovatively, within the LCV2 framework, question and predicted answer pairs are transformed into descriptive and referring captions, enhancing the clarity of the visual cues directed by the question text for the VG module's grounding. This compensates for the limitations of missing intrinsic text–visual coupling in non-end-to-end frameworks. Comprehensive experiments on benchmark datasets, such as GQA, CLEVR, and VizWiz-VQA-Grounding, were conducted to evaluate the method's performance and compare it with several baseline methods. In particular, it achieved an IoU F1 score of 59.6% on the GQA dataset and an IoU F1 score of 37.4% on the CLEVR dataset, surpassing some baseline results and demonstrating the LCV2's competitive performance.

Keywords: large language model (LLM); vision and language; VQA grounding; grounded visual question answering

Citation: Chen, Y.; Su, L.; Chen, L.; Lin, Z. LCV2: A Universal Pretraining-Free Framework for Grounded Visual Question Answering. *Electronics* **2024**, *13*, 2061. https://doi.org/10.3390/electronics13112061

Academic Editors: Arkaitz Zubiaga, Wei Ji, Hao Fei and Fei Li

Received: 7 April 2024
Revised: 13 May 2024
Accepted: 22 May 2024
Published: 25 May 2024

1. Introduction

The task of Grounded Visual Question Answering combines vision and language, evolving from the Visual Question Answering (VQA) task [1]. It requires generating accurate answers based on the input image and question, while accurately grounding the relevant image regions for the question and answer process. Figure 1 provides a visualization of the VQA grounding task. This domain has significant practical relevance, particularly in enhancing visual accessibility. It promises advancements in visual navigation and the precise grounding of objects, offering substantial benefits to individuals with visual impairments [2–5]. Furthermore, certain models and systems [6–9] utilize grounding visual clues or evidence to indicate regions of interest within the visual context, improving the responses' accuracy and interpretability.

The VQA grounding task challenge lies in developing models capable of precisely grounding image regions associated with natural language queries. Existing works [5,7,8,10] typically construct end-to-end deep learning architecture that jointly process visual and linguistic information. These systems utilize classical deep networks [11,12] or attention mechanisms [13,14] to guide the model's focus on areas of the image relevant to the question, achieving modeling of the association between natural language questions and visual objects. However, to understand and integrate multimodal information in an open-world setting, these approaches necessitate extensive cross-modal learning with large samples during the pretraining phase. There is a high computational cost and substantial demand for image–text data during the model training. This limits their modeling in

resource-constrained environments. To address this, we proposed a universal modular framework for VQA grounding tasks. By leveraging the general knowledge of off-the-shelf expert large models, we implement an out-of-the-box, pretraining-free framework. It eliminates the need for any prior large-scale sample or high-cost computational model with pretraining processes, facilitating task applications in resource-limited settings.

Figure 1. The figure shows a visualization of the VQA grounding task, where the system provides textual answers to visual questions and grounds the relevant evidence.

In designing the modular framework for VQA grounding tasks, we conceptualized the VQA grounding task as a deep two-way interaction between the model and visual content. The VQA grounding systems extract detailed visual information to answer questions and provide visual grounding for the relevant regions. Compared to the unidirectional VQA task, which proceeds from visual to textual information, VQA grounding introduces an additional dimension of visual grounding. Based on the above, a prototype for a modular architecture was proposed. We simply combined a general VQA model [15–26] with a general Visual Grounding (VG) model (the VG model in this paper includes Open Vocabulary Object Detection (OVD) models [27–30] and Referring Expression Comprehension (REC) models [31–44]). We utilized the VQA-generated answer or the question–answer pair as the VG model's required grounding captions. This framework is depicted as a naive modular architecture in Figure 2. However, this method encountered two primary issues. Firstly, the non-end-to-end design weakens the intrinsic coupling relationship between the question text and the visual input. This leads to overlooking the visual cues within the question text. Consequently, the model fails to accurately ground the essential visual areas in VQA tasks. Secondly, if the question–answer pair is directly used as the caption for grounding, the lack of contextual referring expression clarity may limit the accuracy of the OVD/REC models in grounding the relevant visual objects.

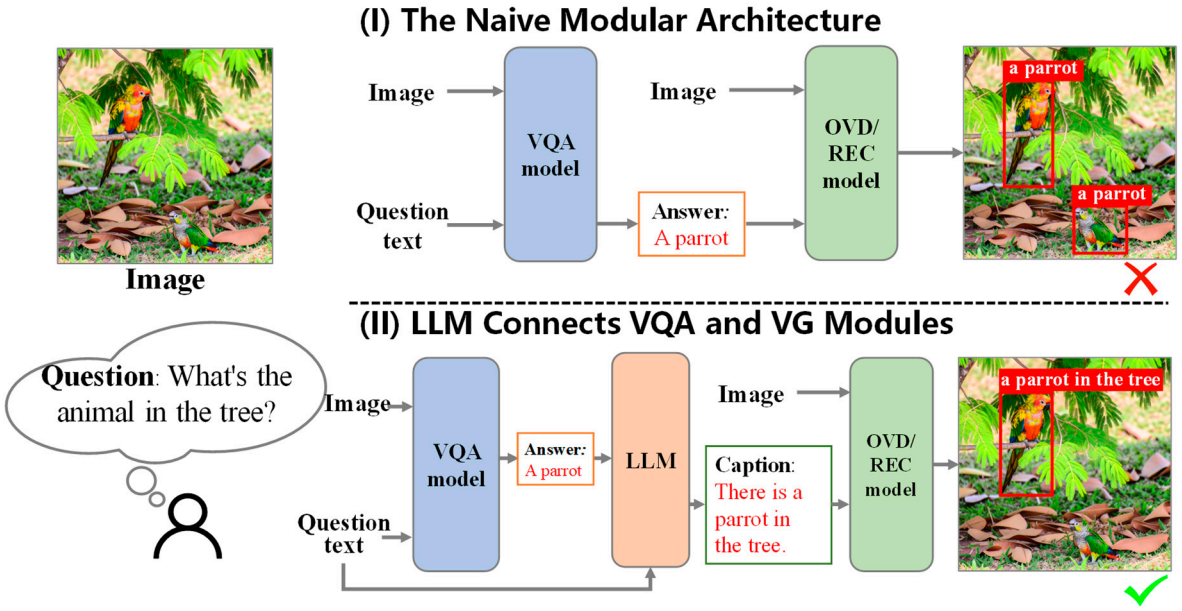


Figure 2. A comparison of two modular frameworks. (**I**) The naive framework fails to use the question text for the visual focus, indiscriminately grounding all parrots, even if the question specifies "the parrot in the tree". (**II**) LCV2, leveraging an LLM, converts the question and answer text into declarative referring captions, enhancing the visual grounding accuracy by focusing on the relevant area in the image.

In response to the issues identified with the above naive modular framework, we proposed a more rational framework design, as shown in part (II) of Figure 2. This framework uses a large language model (LLM) [45–53] to transform the question–answer text pairs—comprising the original question text and the answer text predicted by the VQA model—into highly descriptive referring statements. For instance, the question and predicted answer pair "What is the animal in the tree? A parrot" is reformulated into a more descriptive referring caption: "There is a parrot in the tree". This enhances the text's referring expression and focuses on the visual region corresponding to the question. The caption, along with the original visual content, is then fed into the OVD/REC models for grounding. By designing appropriate prompts to guide the LLM in this text transformation, we have constructed an optimized modular framework system—LCV2 (**L**LM **C**onnects **V**QA and **V**G Modules). This framework cleverly utilizes frozen large language models as a bridge connecting the off-the-shelf VQA module with the off-the-shelf OVD/REC module. It ensures the effective transfer and transformation of the natural language text, fostering tight collaboration between the two modules. This approach, without the need for extensive pretraining, enhances the overall precision of grounding in the VQA grounding task.

Our contributions in this work can be summarized as follows:

(i) This study introduced LCV2, a modular framework specifically designed for the VQA grounding task. It eliminates the need for pretraining, reducing the demand for extensive computational power and data resources.

(ii) LCV2 utilizes a LLM to transform question–answer pair texts into descriptive referring texts suitable for visual grounding. This addresses the issue of insufficient interaction between the visual questioning and grounding in modular frameworks.

(iii) The modules within this universal framework are designed to be plug-and-play and compatible with advanced pretrained models, allowing for dynamic performance improvements as the technology evolves.

(iv) The experimental results demonstrate that our method achieves competitive results on multiple benchmark datasets, including GQA [54], CLEVR [55], and VizWiz-VQA-Grounding [2], compared to other baseline methods.

The upcoming sections of the article are organized as follows:

Section 2, Related Work, introduces the technical background and recent research developments concerning the proposed universal modular framework and VQA grounding task.

Section 3, Methods, delineates the formulation of the problem, further describes the LCV2 framework in the form of mathematical expressions, lists the modular inventory, and finally presents the overall process of framework inference.

Section 4, Experiments, provides information about the public datasets used in the experiments, details of the experimental implementation, evaluation metrics, and a performance comparison of the LCV2 with some baseline models. It also evaluates the VQA module's impact and tests the performance in answering grounding tasks for visually impaired individuals using the Vizwiz test set.

Section 5, Analysis and Discussion, offers a detailed analysis and discussion of the experimental results presented in the paper.

Section 6 discusses the conclusions of this study and the prospects of the proposed framework.

Section 7 elaborates on the limitations of LCV2.

2. Related Work

This section offers a summary of related work pertinent to our method, including developments in VQA and VQA grounding, open vocabulary object detection (OVD), and referring expression comprehension (REC), alongside large language models (LLMs).

2.1. Visual Question Answering and VQA Grounding

Early Visual Question Answering (VQA) efforts focused on adopting joint embedding methods where the extraction of visual features and textual features relied on Convolutional Neural Networks (CNNs) [11] and Recurrent Neural Networks (RNNs) [12] or their variants. The fusion of these cross-modal features was achieved through simple mechanism combinations, such as dot multiplication, dot addition, concatenation, etc. Malinowski et al. [15] combined semantic segmentation with Bayesian methods for the image and question analysis, while Ren et al. [16] enhanced this by feeding encoder outputs to a classifier for answers. Yu et al. [17] and Ben et al. [18] employed multimodal bilinear pooling to integrate spatial image features with question text features.

The introduction of the transformer [13,14] architecture marked a new phase in the cross-modal development. Influenced and inspired by the work of BERT [56] in the field of natural language processing, the visual–text multimodal domain also began to explore the "pretraining–finetuning" paradigm for large models. Numerous pretraining models, based on the transformer architecture, emerged in the visual–text multimodal domain. Examples include ViLBERT [19], VisualBERT [20], Git [21], BLIP [22], BLIP-2 [23], Flamingo [24], Llavar [25], MiniGPT-4 [26], etc. The VQA task was treated as a downstream task in the finetuning training of cross-modal large models and gained important developments and made significant progress in this stage.

As VQA research and cross-modal models advanced, efforts expanded beyond merely enabling VQA systems to answer text questions based on visuals. Some VQA grounding studies [2,5,7,8,57–59] have focused on models that answer questions and ground visual clues simultaneously. The MAC network [58] employs attention-based steps for this, while MAC-Caps [7] enhances the performance with a capsule-based feature selection. Liu et al. [5] further advanced this field with Dual Attention Visual Interaction (DAVI), adding a language-informed visual decoder to the VQA networks. Most recently, the emergence of methods such as xMERTER [60], DDTN [10], and P2G [61] has further advanced the development of VQA grounding techniques. Table A1 presents a summary of these related works. Figure 3 compares the LCV2 framework with established pretrained baselines, MAC-Caps, and DAVI. It also highlights the evolution of the VQA grounding systems. MAC-Caps employs traditional deep learning networks, whereas DAVI adopts the transformer architecture. And now, inspired by the community's shift towards general large

models and modular approach designs, LCV2 offers a universal plug-and-play framework without the need for any retraining.

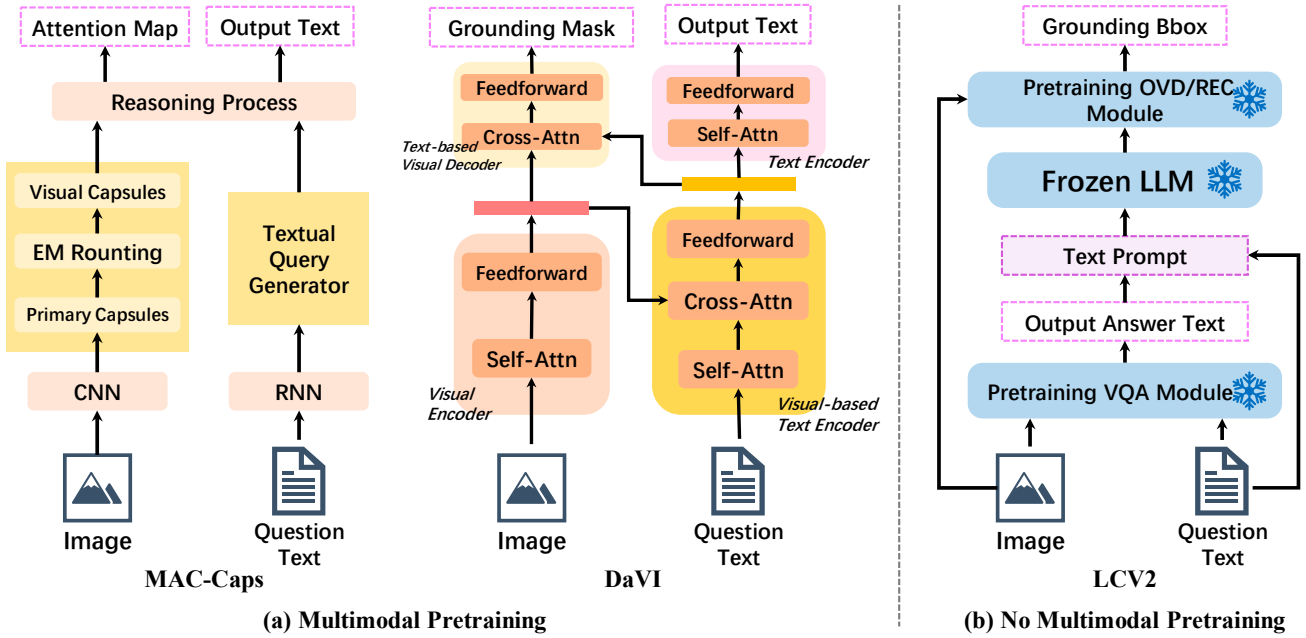


Figure 3. A contrast of our approach with baselines. (**a**) **MAC-Caps** extracts features using a traditional deep neural network; **DaVI** utilizes the transformer-based architecture for encoding and decoding. (**b**) **LCV2**, reflecting the trend towards large, general models with modular designs, integrates a VQA module and an OVD/REC module through an LLM.

2.2. Open-Vocabulary Object Detection and Referring Expression Comprehension

The task of Open-Vocabulary Object Detection (OVD) involves language-based generalization for detecting objects of any category within visual content. In earlier years, OV-DETR [27] utilized a diverse set of image–caption pairs to enhance the model's detection capabilities for unknown classes. Recently, fueled by the transformer architecture [13] and the contrastive learning methods applied in CLIP [28], significant developments have occurred in OVD. GLIP [29] combines object detection with phrase grounding, learning from both data types. DetCLIP [30] introduces a parallel training framework to efficiently utilize diverse datasets.

Referring expression comprehension (REC) involves grounding target areas in an image through natural language referring expressions. Current scholarly work on REC can generally be divided into three categories: (i) structured models, which have evolved from two-stage methods [31–33] to single-stage methods [34–37], like Reclip [36], and which have accelerated the inference speed of models; (ii) multi-task unified models [38–41], which typically combine multiple task modules into a unified framework, such as PolyFormer [41], which processes image–text inputs to predict polygon vertices; and (iii) multimodal pretrained models [42–44], which focus on general visual–language tasks, such as OFA [44], which uses a sequence-to-sequence architecture for various vision–language tasks.

Recent studies highlight the models capable of performing both REC and OVD tasks [62–64]. Grounding DINO [63] proposed a closely integrated solution that enables the model to detect any target in the input text, including category names and referring expressions. Xie et al. [64] expanded this by introducing Described Object Detection (DOD), a combination of OVD and REC. They improved the OFA framework to develop OFA-DOD, which applies to a broader context of object detection based on descriptions.

2.3. Large Language Models

The transformer [13] architecture catalyzed the development of large language models (LLMs). These models were initiated with GPT-1 [65] and BERT [56], establishing the pretraining–finetuning paradigm. Subsequent developments witnessed a continuous increase in the model parameters, giving rise to models such as GPT-2 [66], T5 [67], and OPT [68]. Flan-T5 [45] demonstrated strong generalization capabilities by implementing instruction finetuning on a massively scaled task. The releases of GPT-3 [46] and its successor, GPT-3.5 [47], achieved a remarkable performance across a range of language tasks. GPT-4.0 [48], surpassing GPT-3.5, exhibited significant advancements with cross-modal understanding and generation capabilities, standing as one of the most advanced large models to date. Llama 2 [49], leveraging an optimized autoregressive transformer, showed prowess after pretraining on an extensive token dataset. Other models, such as ERNIE [50], Qwen [51], PaLM [52], and ChatGLM [53], also demonstrated significant capabilities.

This paper summarizes the performance of representative LLMs in Table A2. In constructing the LCV2 framework, the Flan-T5 model was employed as the large language module of the framework. Since the experiments were conducted with relatively limited hardware resources, it was necessary to consider a balance between the model performance and model size.

3. Methods

3.1. Problem Definition

The concrete examples of this task are illustrated in Figure 1. Given an input image $\boldsymbol{X}_I$ and a question text $\boldsymbol{X}_Q$ about the content of the image, the goal of the VQA grounding task is to obtain the correct answer text *Ans* and the grounding result *Grd* of the visual object that the natural language question focuses on. The following sections will introduce the pipeline of our method, which is depicted in Figure 4.

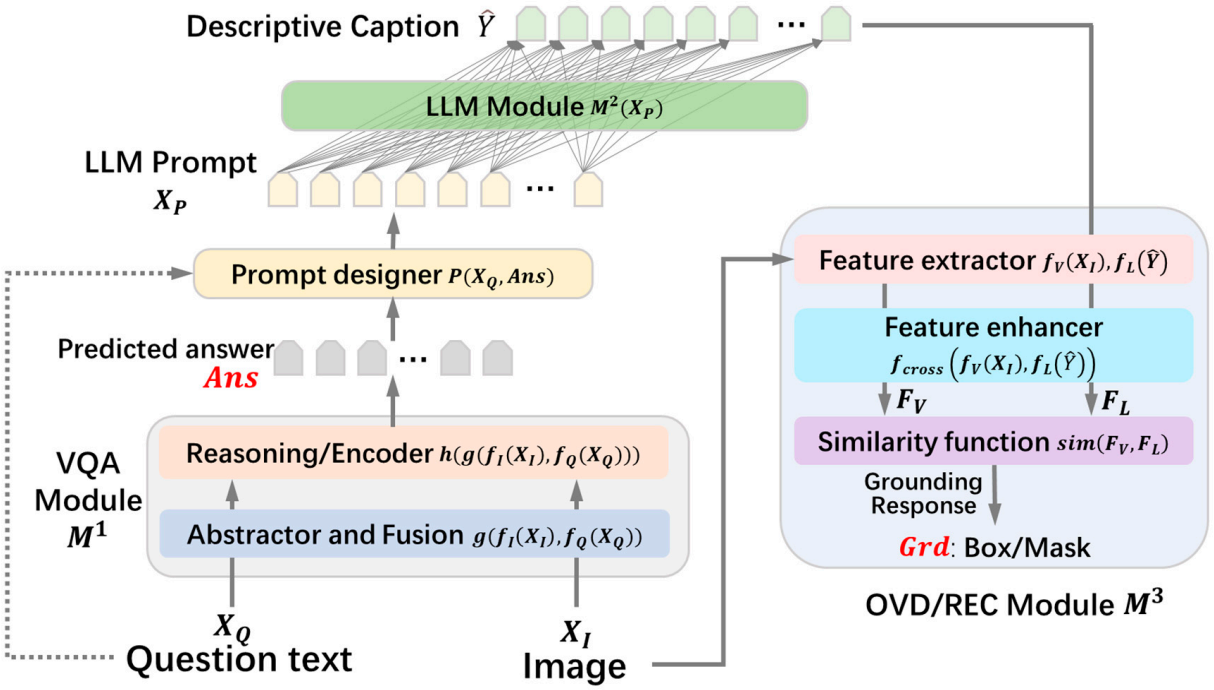


Figure 4. Schematic diagram of the LCV2 modular approach pipeline.

3.2. Modular Framework

Our proposed pretraining-free modular framework consists of a VQA module, an LLM module, and an OVD/REC module. It utilizes an LLM as an intermediate mediator for the conversion of the natural language information to connect the VQA module and the

OVD/REC module. Based on the natural language text representation form between the LLM conversion modules, a unified framework is constructed to complete the VQA grounding task. We name this modular approach LCV2, namely, $LCV2 = (M^i) = (M^1, M^2, M^3)$, where M^i represents the VQA module, LLM module, or OVD/REC module that constitute the framework. The goal of the VQA grounding task can be formalized as Equation (1).

$$Grd,\ Ans = LCV2(\boldsymbol{X}_I, \boldsymbol{X}_Q) \tag{1}$$

$\boldsymbol{X}_I$ and $\boldsymbol{X}_Q$ are first processed and reasoned by the frozen VQA module M^1, generating the predicted answer text Ans, as shown in Equation (2).

$$Ans = M^1(\boldsymbol{X_I}, \boldsymbol{X_Q}) \tag{2}$$

The inference process of the VQA module can be described by four functions, namely, $M^1 = (f_I, f_Q, g, h)$, where $f_I(\cdot)$ and $f_Q(\cdot)$ respectively map the image and question text to vectors; $g(\cdot)$, which is a feature fusion function; and $h(\cdot)$, which is a prediction function, outputting the module's answer text. The process of predicting the answer text Ans by the VQA module can be further detailed as Equation (3).

$$Ans = M^1(\boldsymbol{X_I}, \boldsymbol{X_Q}) = h(g(f_I(\boldsymbol{X_I}), f_Q(\boldsymbol{X_Q}))) \tag{3}$$

The question text $\boldsymbol{X}_Q$ and the predicted answer text Ans are collected, and, based on the text, a prompt for the LLM is designed, as shown in Equation (4). An illustrative example of prompt design is the following: If $\boldsymbol{X}_Q$ is "What is the animal on the tree?" and Ans is "a parrot," then the designed prompt $\boldsymbol{X}_P$ would be "{Question:} What animal is on the tree? {Answer:} A parrot. {Generated content prompt}: convert the above content into a declarative sentence:", which is further fed into the frozen large language model.

$$\boldsymbol{X_P} = P(\boldsymbol{X_Q}, Ans) = \{x_1, x_2, \ldots, x_n\} \tag{4}$$

where x_i is the ith element of the prompt text sequence and $P(\cdot, \cdot)$ is the method for designing the prompt.

Under the guidance of the prompt, the LLM transforms the simple question and predicted answer text ("What is the animal in the tree? A parrot".) into a declarative caption description ("There is a parrot in the tree".), making the question and answer text contextually fluent and referring. This transformation enables visual grounding tasks to accurately ground the visual region that is referenced by the question and the predicted answer, based on the descriptive caption. The inference process of the LLM module M^2 can be represented by four functions, namely, $M^2 = (f_H, f_O, sampling, P)$, where f_H is used to compute the hidden representations; f_O outputs the probability distributions; and sampling defines a sampling function based on the given probability distribution. The hidden representation by LLM is shown as Equation (5):

$$\boldsymbol{H} = f_H(\boldsymbol{X_P}, \boldsymbol{\theta}) = f_H(\{x_1, x_2, \ldots, x_n\}, \boldsymbol{\theta}) \tag{5}$$

where $\boldsymbol{\theta}$ represents the model parameters and the aforementioned $f_H(\cdot)$ and $f_O(\cdot)$ are the functions defined based on $\boldsymbol{\theta}$. Based on the hidden representation, LLM outputs a probability distribution as shown in Equation (6).

$$\boldsymbol{Y_i} = f_o(\boldsymbol{H}, \boldsymbol{\theta} | \{\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y_{i-1}}\}) \tag{6}$$

where $\boldsymbol{Y}_i$ represents the output probability at time step i. Based on the output probability distribution, the sampling generation process by LLM is as follows:

$$\hat{y}_i \sim sampling(\boldsymbol{Y_i}) \tag{7}$$

where $\hat{\boldsymbol{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_i, \ldots, \hat{y}_n)$ represents the text sequence generated by the LLM, which is the descriptive captions suitable for the visual grounding. Finally, $\hat{\boldsymbol{Y}}$ and $\boldsymbol{X}_I$ are further fed into the frozen OVD/REC module. The OVD/REC module associates the captions $\hat{\boldsymbol{Y}}$ with objects, scenes, or regions in the $\boldsymbol{X}_I$, predicting the grounding result, i.e., $Grd = M^3(\boldsymbol{X}_I, \hat{\boldsymbol{Y}})$, where M^3 represents the inference process of the OVD/REC module and $M^3 = (f_V, f_L, f_{cross}, sim)$, consisting of f_V, the visual feature extraction function; f_L, the textual language feature extraction function; f_{cross}, the feature fusion enhancement function; and sim, the similarity calculation function. Equation (8) describes the process of the visual and language feature extraction and feature enhancement:

$$\boldsymbol{F}_V, \boldsymbol{F}_L = f_{cross}\left(f_V(\boldsymbol{X}_I), f_L\left(\hat{\boldsymbol{Y}}\right)\right) \tag{8}$$

The similarity between the visual and language features is calculated, and the region or object in image $\boldsymbol{X_I}$ that best matches the description $\hat{\boldsymbol{Y}}$ is determined, as follows:

$$Grd = sim(\boldsymbol{F}_V, \boldsymbol{F}_L) = Argmax_{\boldsymbol{region}\ in\ \boldsymbol{X}_I} sim\left(\boldsymbol{region}, \hat{\boldsymbol{Y}}\right) \tag{9}$$

Therefore, summarizing the above, the overall LCV2 unified architecture can be mathematically formalized as Equation (10).

$$Ans, Grd = LCV2(\boldsymbol{X}_I, \boldsymbol{X}_Q) = M^3\left(M^2\left(P\left(M^1(\boldsymbol{X}_I, \boldsymbol{X}_Q), \boldsymbol{X}_Q\right)\right), \boldsymbol{X}_I\right) \tag{10}$$

In the sections below, the key points of each module will be introduced, outlining the main modular components involved in the framework designed in this paper.

3.3. Modular Inventory

This section introduces the three primary modular models that constitute the LCV2, including the VQA module, LLM module, and OVD/REC module. Each of these modules leverages off-the-shelf general purpose expert models to accomplish their respective downstream tasks.

VQA module. We focused on exploring different model alternatives for the VQA module, specifically including BLIP-VQA, Lens, and Git-VQA. During the model selection process, we focused on lightweight models for fast inference with limited computational and storage resources. **BLIP-VQA** [22] contains a Vision Transformer (ViT) [14] image encoder, an image-guided text encoder, and a decoder. The model's performance is enhanced through image–text contrast (ITC), image–text matching (ITM), and language modeling (LM) pretraining objectives. **Lens** [69] is a pretraining-free VQA modular approach proposed by Stanford University, which combines visual representation models, such as CLIP or BLIP, to generate exhaustive labels, attributes, and captions and uses a large language model for inference to produce the final answer. **Git-VQA** [21] excels in conventional visual question answering and is also adept at identifying the text in images. Based on a straightforward architecture, the image is processed through a visual encoder and, together with the text's tokenization and embedding, enters a text decoder with multi-head attention and feed-forward layers.

LLM Module. The large language model component was constructed as a flexible and replaceable unit that could conveniently adopt different state-of-the-art (SOTA) LLMs, such as the ChatGPT series [46–48,65,66], the LLaMa series [49], ChatGLM [53], or others. In the experimental implementation of this research, we selected the frozen Flan-T5-large [45] as the LLM module of the framework. As shown in Table A2, which presents the performance and model parameters, we primarily considered the balance between lightweight parameters and well performance across various NLP tasks.

OVD/REC Module. The OVD/REC module was implemented based on the Grounding DINO [63] model, which possesses the dual capabilities of open-vocabulary object detection and referring expression comprehension. In our experiments, we evaluated and applied two versions of Grounding DINO: Grounding DINO-T and Grounding DINO-B.

The former is pre-rained on datasets such as O365, GoldG, and Cap4M, while the latter is pretrained on a broader set of datasets, including COCO, O365, GoldG, Cap4M, OpenImage, etc. Both versions of the model demonstrate an outstanding performance on open-world OVD/REC tasks.

3.4. Framework Inference

The VQA, LLM, and OVD/REC modules process and infer the data in a sequential manner. They use the output results from each module as the input for the next. The expert models that make up the LCV2 are arranged in a streaming architecture. Initially, the visual content and question text are inputted into the pretrained VQA module to generate a preliminary predicted answer. The original question text and this predictive answer text are then collected to design a prompt that directs the work of the LLM. To be more precise, the standard format of this prompt is designed as follows: "{questions:} 'question text' {answer:} 'response text' {generated content prompt} 'convert the above content into a declarative sentence:'". Subsequently, the LLM module is applied. The comprehensive prompt guides the LLM in generating the text content that fits the captions required for the grounding by the OVD/REC model. Finally, the descriptive caption is then inputted into the OVD/REC module to achieve the precise target grounding and referring expression comprehension.

4. Experiments

4.1. Datasets

This research validated the performance of the LCV2 on the GQA and CLEVR validation sets, comparing it with baseline methods, and assessed its effectiveness in answer grounding for visually impaired individuals using the VizWiz-Answering-Grounding dataset against SOTA methods. This section summarizes the datasets used in our experiments.

GQA. The GQA dataset [54,70], created by Stanford University's Manning group, aims to improve real-world visual reasoning in VQA tasks by reducing the language bias found in previous datasets. Unlike VQA 2.0 [71], GQA focuses more on testing the models' reasoning and compositional skills. It includes scene graphs for each image, detailing objects, their properties, and spatial information, enhancing question realism and aiding in VQA grounding task training. The experiment evaluates the method's performance on the GQA balanced version of the validation set, which provides a total of 132,062 question and answer pairs along with other necessary information.

CLEVR. The CLEVR dataset [55] by Johnson et al. [55] is a synthetic dataset designed for evaluating visual–textual reasoning in 'VQA tasks. It features abstract geometric shapes instead of real-world images to focus on visual reasoning with minimal bias. The questions are programmatically generated, covering areas like counting, comparison, existence, and attributes (color, material, and size), to test a broad spectrum of reasoning skills. This experiment evaluates the method on the CLEVR v1.0 validation set, involving the finetuning of the VQA module using the training set provided by the dataset.

VizWiz-Answering-Grounding. The VizWiz dataset [2,72] supports VQA tasks for visually impaired users, sourced directly from this community rather than through crowdsourcing, ensuring relevance to their real-world needs. The VizWiz-VQA-Grounding subset includes annotations for object grounding, aiding in model training and evaluation. It contains about 6.4 k training, 1.1 k validation, and 2.3 k test examples. The experimental evaluation of the method presented in this paper was conducted on the test set of VizWiz-Answering-Grounding. The finetuning of the VQA module for this dataset was completed using the training set provided by the dataset.

4.2. Implementation Details

The experimental implementation of LCV2 used off-the-shelf VQA modules like BLIP-VQA-Capfilt-large, Lens, and Git-large-VQAv2, sourced from Hugging Face. Specifically, BLIP-VQA-Capfilt-large features a large ViT backbone trained on the VQA2.0 dataset. Lens

employs CLIP-H/142 and CLIP-L/143 for tagging and attributing visual content and uses BLIP-Image-Captioning-large for generating subtitles, while its reasoning tasks are handled by the Flan-T5-large LLM. Git-large-VQAv2, a larger model in the Git series, is finetuned on the VQA v2 dataset. The LLM module uses Flan-T5-large with 750 million parameters. The OVD/REC module of the LCV2 utilizes the Grounding DINO model with REC and OVD capability, where we examine the impact of both swin-B and swin-T versions of the model on the experimental results.

Sections 4.5 and 4.6 detail the finetuning of the VQA module using the Low-Rank Adaptation (LoRA) method. The transformer architecture employs a well-known self-attention mechanism defined as $\mathrm{Att}(\boldsymbol{X}_q\boldsymbol{W}_Q, \boldsymbol{X}_k\boldsymbol{W}_K, \boldsymbol{X}_v\boldsymbol{W}_V) = \mathrm{softmax}\left(\frac{\boldsymbol{X}_q\boldsymbol{W}_Q \cdot (\boldsymbol{X}_k\boldsymbol{W}_K)^T}{\sqrt{d}}\right)\boldsymbol{X}_v\boldsymbol{W}_V$, with $\{\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V\}$ as the linear transformation matrices and $\{\boldsymbol{X}_q, \boldsymbol{X}_k, \boldsymbol{X}_v\}$ as the input parameters. LoRA introduces low-rank matrices $\boldsymbol{W}_A$ and $\boldsymbol{W}_B$ as bypass matrices to update the VQA model, formulated as $\boldsymbol{Y} = \boldsymbol{W}_{pre}\boldsymbol{X} + \boldsymbol{W}_{tr}\boldsymbol{X} = \boldsymbol{W}_{pre}\boldsymbol{X} + \boldsymbol{W}_B\boldsymbol{W}_A\boldsymbol{X}$, where $\boldsymbol{W}_A \in \mathbb{R}^{\alpha\times\beta}$, $\boldsymbol{W}_B \in \mathbb{R}^{\beta\times\gamma}$ and $\beta \ll \min(\alpha, \beta)$, with $\boldsymbol{W}_{tr}$ representing the low-rank matrix added to $\boldsymbol{W}_Q$ and $\boldsymbol{W}_V$ in the transformer architecture during the finetuning. Further hyperparameter details are provided in the corresponding experimental subsections.

We conducted our experiments using the PyTorch framework version 2.1.2 with CUDA version 11.8 on an Ubuntu 18.04.6 LTS 64-bit operating system. To assess the feasibility and effectiveness of our modular approach on lower computational and memory resources, we implemented the framework relying on a set of lower performance hardware resources for the inference and computations. The CPU model used was the Intel(R) Xeon(R) Silver 4210R, and the GPU selected was the Nvidia GeForce RTX 2080Ti, with a VRAM space of only 11 GB.

4.3. Evaluation Metrics

We assessed the VQA task performance of the models based on the accuracy of their generated responses. To evaluate the performance of our framework and baseline methods in VQA grounding for the VQA tasks on the CLEVR and GQA datasets, we relied on precision (P), recall (R), and F1 score metrics to report the Intersection Over Union (IoU) and Overlap metrics for the answer grounding. However, for the VizWiz-VQA-Grounding dataset, the ground truth for the answer grounding is presented in the form of image segmentation based on all points along the object boundaries. Therefore, the evaluated models, after obtaining the grounding results, need to further process the predicted grounding regions and convert them into binary mask images. The IoU metric is then calculated using the pixel counts of the intersection and the union of the masks.

The computation of the IoU metric can be mathematically described as follows:

$$IoU = \frac{Area(\boldsymbol{preRGN} \cap \boldsymbol{gtRGN})}{Area(\boldsymbol{preRGN} \cup \boldsymbol{gtRGN})} \tag{11}$$

where $Area(\cdot)$ denotes the method for calculating the area of a region, ***preRGN*** represents the predicted region, and ***gtRGN*** stands for the ground truth region label.

Equation (12) gives the mathematical expression of the overlap metric:

$$Overlap = \frac{Area(\boldsymbol{preRGN} \cap \boldsymbol{gtRGN})}{Area(\boldsymbol{gtRGN})} \tag{12}$$

In experiments, a specific IoU threshold is set to further calculate precision (P), recall (R), and the F1 score, reporting the answer grounding performance of the evaluated methods.

4.4. Comparison with Baseline Models

We assessed the VQA accuracy and grounding performance of the LCV2 framework against baseline models on the CLEVR and GQA datasets. The baselines included MAC, MAC-CAP, SNMN, and SNMN-Caps. These baseline models processed the visual and

textual inputs to generate attention maps, which were analyzed using connected component analysis to create bounding boxes for the unified grounding evaluation. The LCV2 was tested in two versions, differentiated by the Grounding DINO model in the OVD/REC module (swin-B or swin-T). The Grounding DINO hyperparameters were set with box and text thresholds at 0.25.

GQA Setting. The balanced version of the GQA validation set includes ground truth bounding box labels for phrase grounding associated with question texts, answer texts, full answer texts, and all texts. The LCV2 and baseline methods were evaluated using the "answer" and "all" text grounding labels. For the "answer" text evaluations, the LCV2's predicted answers were directly used in the VG module to predict the grounding Bbox results without LLM processing, reflecting the labels' focus on "answer" text grounding. Figures 5 and 6 partially showcase the visual results of the LCV2 on the GQA dataset. Table 1 details performance comparisons between the LCV2 and baseline methods, including IoU and Overlap metrics with a set IoU threshold of 0.5. The baseline models MAC and MAC-Caps have a time step T set to 4, as they have been reported to perform the best on this dataset.

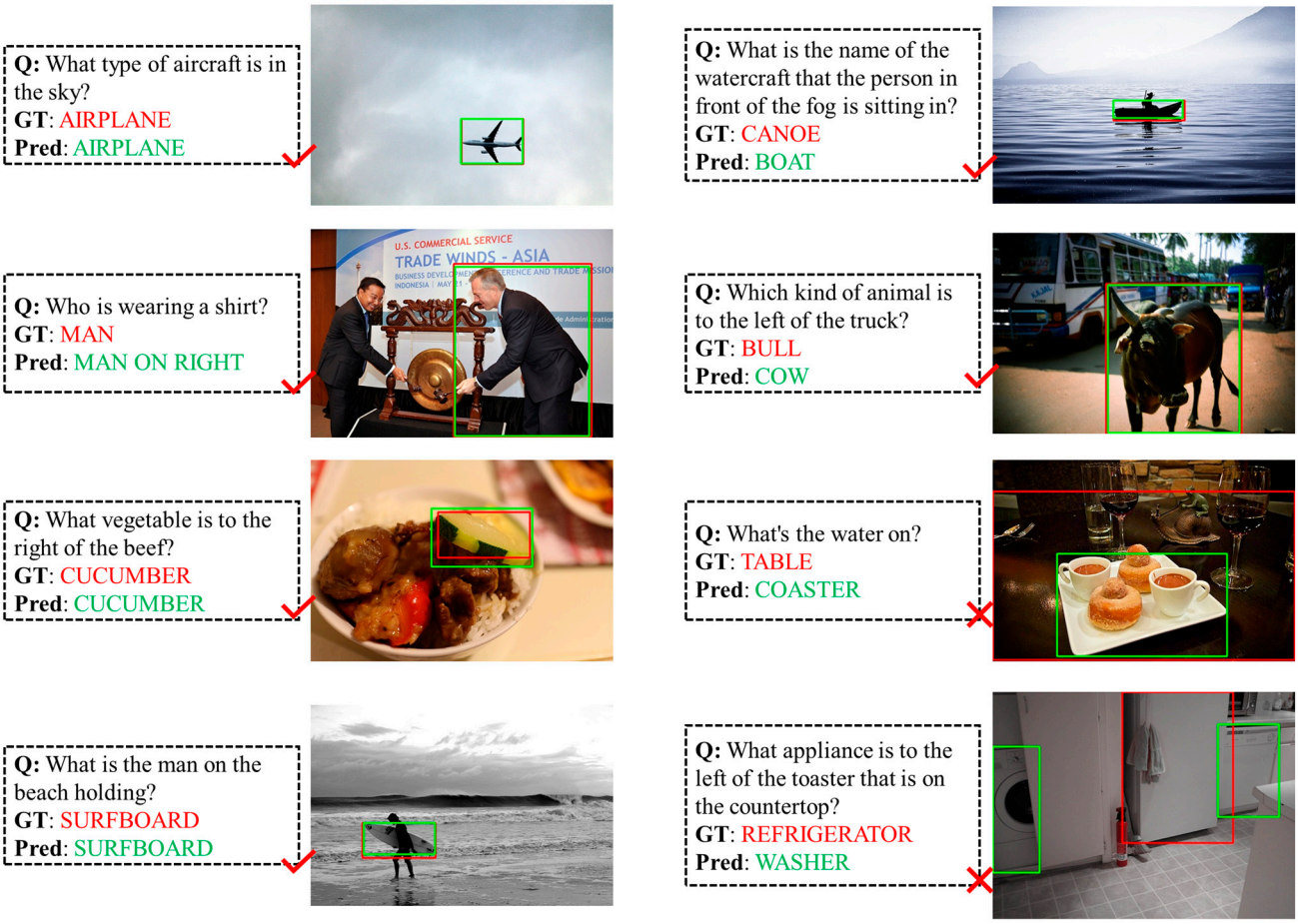


Figure 5. Visualization examples from the LCV2 model on the GQA balanced version's validation set show the grounding results for the "answer" text. The red bounding boxes mark the ground truth object and the green boxes show the LCV2's predictions.

Result. The LCV2, based on the BLIP-VQA-Large VQA module without further fine-tuning, achieved an answer accuracy of 56.6%, which represents a 1.5% improvement over MAC-Caps. The LCV2 demonstrated significant performance gains in visual cue grounding, as evidenced by the F1 scores for IOU and Overlap. The visualization comparison in Figure 7 also demonstrates the accuracy of the LCV2 in visual object grounding. The LCV2, incorporating the Grounding DINO Swin-B version in its OVD/REC module, achieved the

best performance in phrase grounding on answer texts, with an IoU F1 score and overlap F1 score of 0.417 and 0.614. Meanwhile, the LCV2, using the Grounding DINO Swin-T version in its OVD/REC module, achieved the best performance in phrase grounding for all texts. In Section 5.1 below, we discuss the reasons for the LCV2's significantly superior F1 score compared to baseline methods, in conjunction with the visualization results from Figure 7, and the advantages demonstrated by the LCV2.

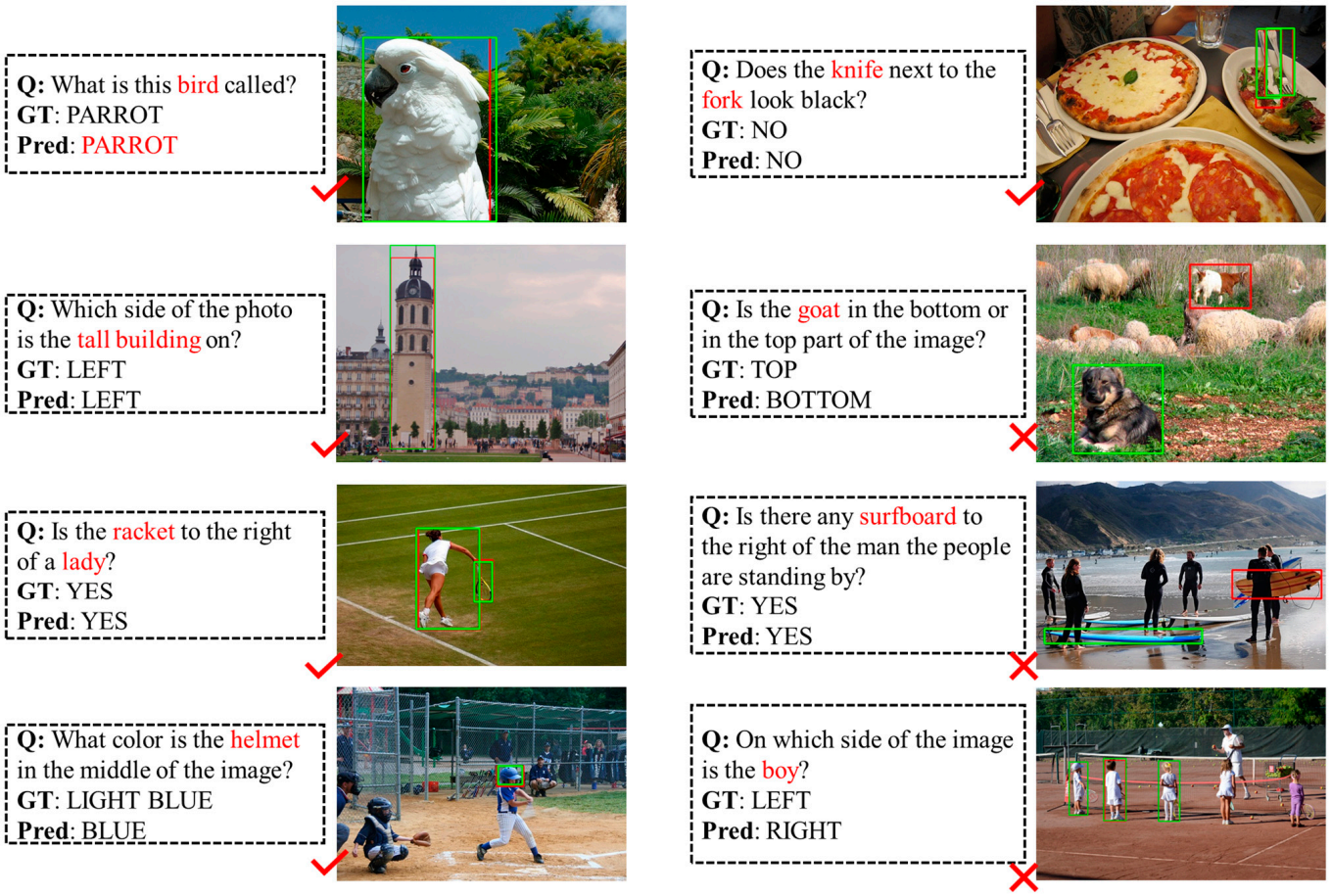


Figure 6. Visualization examples from the LCV2 model on the GQA balanced version's validation set for "all" text. The red bounding boxes represent the ground truth labels, while green bounding boxes depict the bounding boxes predicted by LCV2. Best viewed in color.

Table 1. Experimental validation of the LCV2 on the GQA validation set, compared with baselines. The table shows two LCV2 versions, LCV2 (swin-T) and LCV2 (swin-B), differentiated by their VG modules: Grounding DINO swin-T or swin-B.

Models	Obj.	Acc.	IoU			Overlap		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
MAC [58]	A	**0.571**	0.009	0.045	0.015	0.056	0.274	0.093
MAC-Caps [7]		0.551	0.023	0.119	0.039	0.120	0.626	0.201
LCV2 (swin-T)		0.566	0.273	**0.637**	0.382	0.372	0.786	0.505
LCV2 (swin-B)		0.566	**0.323**	0.590	**0.417**	**0.497**	**0.805**	**0.614**
MAC [58]	All	**0.571**	0.037	0.043	0.040	0.250	0.305	0.275
MAC-Caps [7]		0.551	0.070	0.087	0.078	0.461	0.623	0.530
LCV2 (swin-T)		0.566	0.515	**0.707**	**0.596**	0.751	**0.894**	**0.816**
LCV2 (swin-B)		0.566	**0.516**	0.659	0.578	**0.763**	0.856	0.807

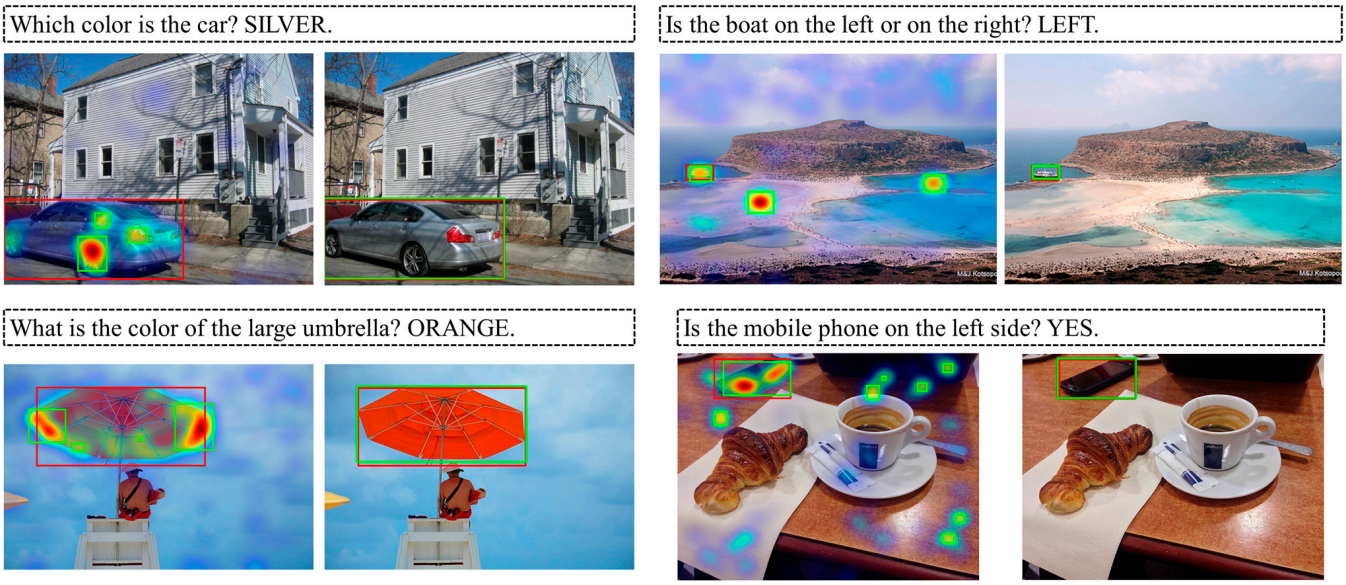


Figure 7. Comparison of the visual grounding results between baseline methods and the LCV2. The baselines extract visual object bounding box positions based on connected components from attention maps. The predicted bounding boxes are shown in green, while the ground truth bounding boxes are shown in red.

CLEVR Setting. The CLEVR validation set used to evaluate the method includes 149,991 question and answer pairs, along with explicitly referenced bounding box labels for the questions and answers. For the baseline models MAC and MAC-Caps, we set the model's time step T to 4, 6, and 12 to compare the performance of different configurations of the baseline models. For the baseline models SNMN and SNMN-Caps, T was set to 9, as it has been reported to show a good performance with this parameter value. We visualized the results of the LCV2's inference on the CLEVR validation set, as shown in Figure 8. The quantitative results of the experiments are presented in Table 2, where the experiment data were obtained with an IoU threshold set to 0.5.

Table 2. Experiments were conducted on the CLEVR validation set, and the performance was quantitatively compared with baseline methods.

Models	T	Acc.	IoU			Overlap		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
MAC [58]	4	0.977	0.140	0.335	0.197	0.249	0.563	0.346
MAC-Caps [7]		0.968	0.240	0.391	0.297	0.470	0.731	0.572
MAC [58]	6	0.980	0.126	0.236	0.164	0.301	0.524	0.382
MAC-Caps [7]		0.980	0.290	0.476	0.361	0.485	0.798	0.603
MAC [58]	12	**0.985**	0.085	0.181	0.116	0.287	0.533	0.373
MAC-Caps [7]		0.979	0.277	0.498	0.356	0.509	0.946	0.662
SNMN [73]	9	0.962	0.378	0.475	0.421	0.529	0.670	0.591
SNMN-Caps [7]		0.967	**0.506**	0.518	**0.512**	**0.738**	0.781	**0.759**
LCV2 (swin-T)	-	0.367	0.265	**0.577**	0.363	0.418	**0.785**	0.545
LCV2 (swin-B)		0.367	0.296	0.425	0.349	0.492	0.660	0.564

Result. The LCV2, utilizing the BLIP-VQA-Large VQA module, achieved a suboptimal answer accuracy of only 36.7% on the CLEVR dataset without any further model training specific to the CLEVR scenario. Additionally, the LCV2's performance in visual grounding, as measured by IoU and overlap F1 score metrics, lagged behind SNMN and SNMN-Caps, with SNMN-Caps achieving the best grounding performance in the experiment.

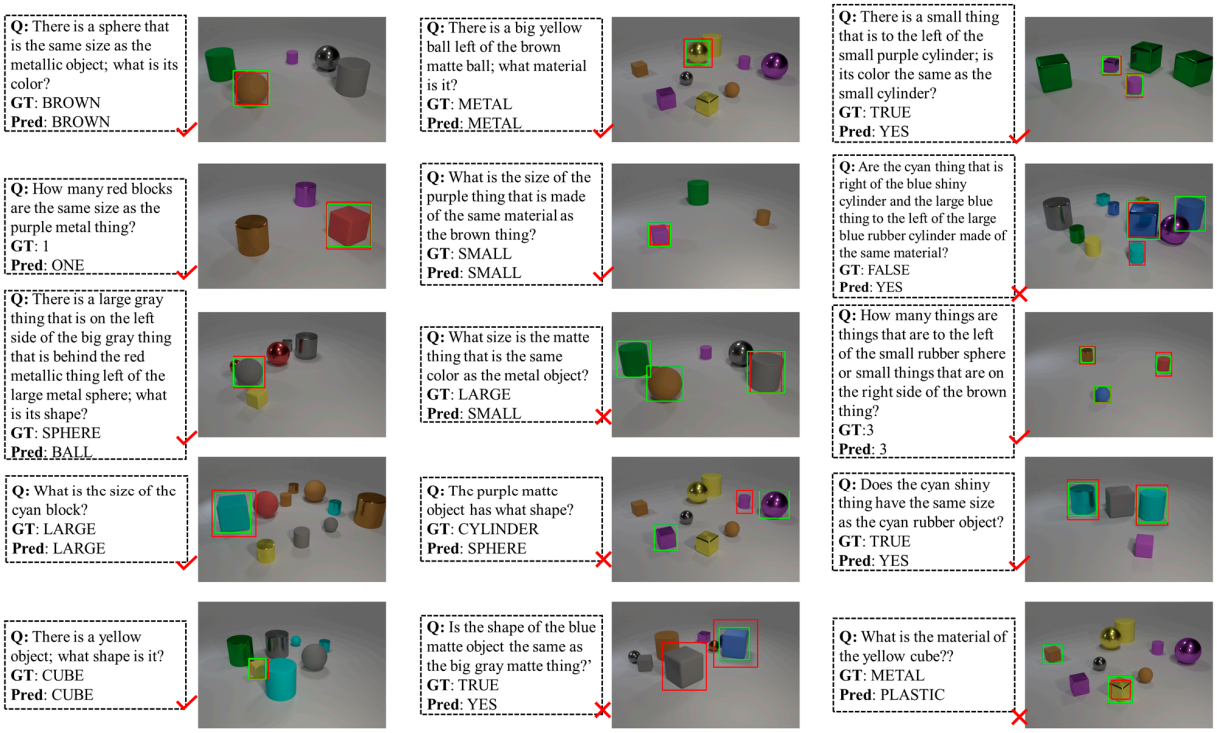


Figure 8. Visualization examples of the LCV2's VQA grounding on the CLEVR validation set display predicted and ground truth answers. The green bounding boxes show the LCV2's predictions, while the red boxes indicate the ground truth.

4.5. Impact of the VQA Modules on Results

The VQA module is a critical component in the LCV2, providing crucial predicted answers to the visual questions. The performance of the VQA module significantly impacts the grounding performance of the LCV2. Therefore, this section conducts experiments to test the impact on the task performance of various pretrained VQA models or frameworks as the VQA module in the LCV2 framework. Consistent with Section 4.4, the LLM module uses the frozen Flan-T5-large model. The experiments are conducted on the same datasets as in Section 4.4.

GQA Setting. On the GQA validation set, we employed publicly available off-the-shelf models, namely, Lens, BLIP-VQA-Capfilt-large, and Git-large-VQAv2, to serve as the VQA module in the LCV2. The LLM and OVD/REC modules were not replaced, resulting in three versions of the LCV2 framework. The OVD/REC module of the LCV2 was constructed using the Grounding DINO swin-B version, which was pretrained on a more extensive dataset, theoretically enabling more accurate grounding of open-world objects. The experiments were conducted based on the ground truth labels for phrase grounding using the "answer" text and "all" text from the dataset. The performance changes reported by LCV2 based on different off-the-shelf VQA systems were evaluated using the IoU and Overlap metrics. The experimental data are presented in Table 3, and the experimental data were obtained with an IoU threshold set at 0.5.

Result. The LCV2's VQA module based on BLIP-VQA-Capfilt-large achieved the highest answer accuracy at 56.6%, and its performance metrics for the visual cue grounding were also the best. In comparison, the LCV2 equipped with the Lens-based VQA module exhibited the poorest performance in both visual question answering and visual cue grounding. We visualized the inference results, as shown in Figure 9. This study also evaluated the computational load and parameter size of the three LCV2 versions using FLOPs to measure complexity during inference. Table 4 details these metrics for the LCV2's components and versions. The BLIP-VQA-Capfilt-large VQA module has a lower computational load and smaller parameter size yet delivers the best performance. In

contrast, Lens has the highest load and size but the weakest performance. The LCV2 (Blip-L) version provides a better balance between the model's complexity and performance.

Table 3. Validation of the LCV2's impact on the GQA validation set assessed three versions with different VQA modules. The experimental results focused on the grounding referenced objects in the Answer (A) and All texts.

Models	Obj.	Acc.	IoU			Overlap		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
LCV2 (Blip-L)	A	**0.566**	**0.323**	**0.590**	**0.417**	**0.497**	**0.805**	**0.614**
LCV2 (lens)		0.278	0.261	0.505	0.345	0.491	0.795	0.607
LCV2 (Git)		0.518	0.292	0.545	0.380	0.463	0.770	0.579
LCV2 (Blip-L)	All	**0.566**	**0.516**	**0.659**	**0.578**	**0.763**	0.856	**0.807**
LCV2 (lens)		0.278	0.414	0.612	0.494	0.692	**0.858**	0.776
LCV2 (Git)		0.518	0.506	0.649	0.568	0.756	0.852	0.801

Figure 9. Examples of the VQA grounding results for the LCV2 on the GQA balanced validation set are shown. Each column represents the predictions made by the LCV2 based on a specific VQA model. The green and red bounding boxes represent the predicted and ground truth bounding boxes, respectively.

CLEVR Setting. Since the visual scenes in CLEVR are not captured from the real world but are composed of abstract three-dimensional geometric shapes generated by computers, the publicly available off-the-shelf BLIP-VQA-Capfilt-large model, pretrained on real-world visual content, exhibits a relatively modest performance on the CLEVR

VQA task. To enhance the model's performance, we conducted efficient finetuning of the BLIP-VQA-Capfilt-large model on the CLEVR training set using the LoRA method. The CLEVR training set provides 70 k images and 699,989 pairs of question–answer text. For efficient finetuning, trainable bypasses were added to the Q-matrix and V-matrix of the transformer framework in BLIP-VQA-Capfilt-large. The rank of the trained matrices was set to 16, the scaling factor to 16, and the dropout ratio to 0.1; no bias parameters were trained. The number of parameters retrained in the BLIP-VQA-Capfilt-large model amounted to 2,359,296, constituting 0.610% of the total model parameters. Considering the available hardware resources, the training time for one epoch was approximately 8 h, and we conducted a simple five-epoch finetuning to obtain the BLIP-VQA-Large-finetuned model on CLEVR. The experimental results are presented in Table 5. To further validate the generalization performance of the LCV2 and assess its stability when confronted with different types of visual questions, we evaluated the LCV2 on CLEVR validation sets divided by problem type according to counting, existence, number comparison, attribute (color, material, and size) comparison, and attribute query. The performance results were measured using the IoU and Overlap F1 score metrics, and the results are visualized in Figure 10.

Table 4. The computational complexity and parameter quantity of the evaluated modules and frameworks.

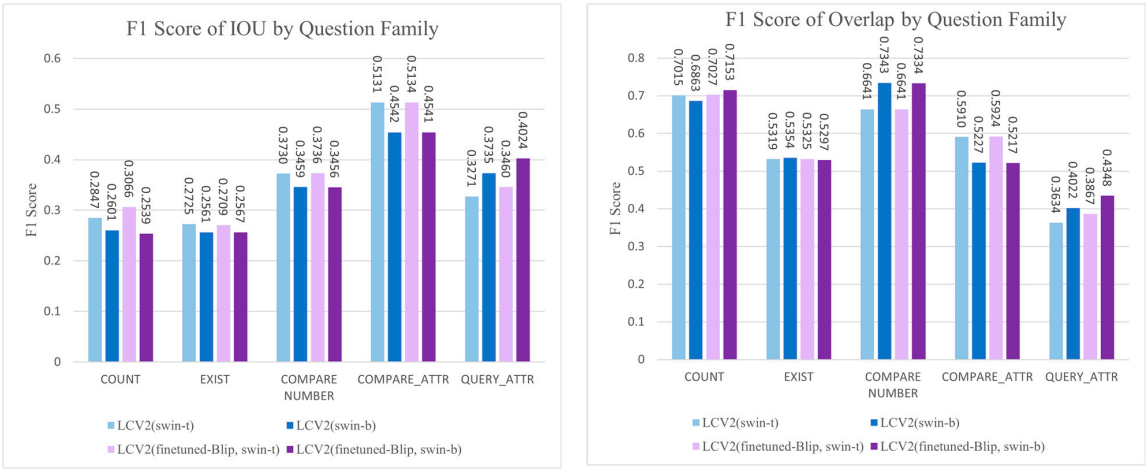
Models	FLOPs	Params
BLIP-VQA-Capfilt-large	59.063 G	336.557 M
Git-large-VQAv2	333.409 G	369.705 M
Lens	3377.488 G	1077.052 M
Flan-T5-large	13.075 G	750.125 M
Grounding DINO swin-T	39.177 G	144.140 M
Grounding DINO swin-B	58.812 G	204.028 M
LCV2 (Blip-L)	130.950 G	1290.710 M
LCV2 (lens)	3449.375 G	2031.205 M
LCV2 (Git)	405.296 G	1323.858 M

Table 5. The experiment mainly validates the impact of two versions of the BLIP-VQA-Capfilt-large model, one finetuned using LoRA on the CLVER training set and the other without finetuning.

Models	Acc.	IoU			Overlap		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
LCV2 (swin-T)	0.367	0.265	0.577	0.363	0.418	0.785	0.545
LCV2 (swin-B)	0.367	0.296	0.425	0.349	0.492	0.660	0.564
LCV2 (finetuned-Blip, swin-T)	**0.773**	0.273	**0.596**	**0.374**	0.424	**0.801**	0.554
LCV2 (finetuned-Blip, swin-B)	**0.773**	**0.312**	0.425	0.360	**0.512**	0.662	**0.577**

Result. Using the BLIP-VQA-Capfilt-large model finetuned on the CLEVR dataset in the LCV2 framework significantly increased the answer accuracy to 0.773, which also improved the visual content grounding accuracy. The LCV2 with Grounding DINO Swin-T achieved the highest IoU F1 score at 0.374, while the version with Grounding DINO Swin-B recorded the best Overlap F1 score at 0.577. In validating the generalization performance of the LCV2, we primarily analyzed the visualization results provided by the IoU F1 score metric, because the IoU metric is more comprehensive, considering the overall match between the predicted and actual areas. The LCV2 demonstrated a balanced performance across multiple types of visual question answering grounding, confirming the model's stability across different types of questions on the CLEVR dataset. The results also showed

that the method performed more distinctly in attribute comparison relative to other types of questions, followed by number comparison and attribute query types.



(**a**) (**b**)

Figure 10. Performance of the method on different types of questions in the CLEVR dataset in terms of IoU and overlap F1 scores. (**a**) The F1 score of IoU for LCV2 on different types of questions. (**b**) The F1 score of overlap for LCV2 on different types of questions.

4.6. LCV2 on the VizWiz Answer Grounding

A practically significant application of the VQA grounding task is to assist visually impaired individuals in answering questions about visuals and to provide explicit grounding for the referenced objects within the image. The VizWiz Answer Grounding dataset benchmarks the performance of methods in answer grounding tasks for visually impaired individuals.

In this section, we deploy the LCV2 on the test set of the VizWiz Answer Grounding dataset, which comprises 2373 test examples. We actively participated in and submitted entries for the online challenge hosted by EvalAI: test-standard2023-vizwiz-VQA-Grounding. The evaluation results are reported using the average IoU metric. We compare the LCV2 with some SOTA methods, and the experimental data are presented in Table 6. The evaluated LCV2 framework is composed of BLIP-VQA-Capfilt-large finetuned on the VizWiz Answer Grounding training set, the large language model Flan-T5, and the Grounding DINO swin-B. The results of the VQA grounding tasks inferred by LCV2 are visualized in Figure 11. The answer grounding results are transformed into binary mask images, where the foreground or the internal region of the grounding is represented by white and the background or the external region is represented by black.

Table 6. An evaluation of the performance of the LCV2 on the VizWiz Answer Grounding test set and a comparison with some state-of-the-art (SOTA) methods, with the results reported in terms of the IoU metric.

Year	Team	IoU
2022	Aurora (ByteDance and Tianjin University)	0.71
	hsslab_inspur	0.70
	Pinkiepie	0.33
	binggo	0.08
2023	UD VIMS Lab (EAB)	0.74
	MGTV_Baseline	0.72
	DeepBlue_AI	0.69
	USTC	0.46
	ours	0.43

Figure 11. Visualization of several prediction results of the LCV2 on the VizWiz Answer Grounding test set. The LCV2's predicted answer grounding is presented in the form of binary mask images.

To investigate the impact of different VQA models on the task performance of the LCV2, comparative experiments were conducted. We employed Lens, GittextVQA, BLIP-VQA-Capfilt-large model, and the finetuned BLIP-VQA-Capfilt-large model as the VQA module of the LCV2, observing the influences on the final answer grounding's IoU quantification results. It is worth mentioning that the utilized GittextVQA is based on the Git-large-VQAv2 model finetuned on the TextVQA dataset, possessing the ability to recognize text in visuals. The experimental data are presented in Table 7. The LCV2 constructed with BLIP-VQA-Capfilt-large and Grounding DINO swin-B achieved an IoU of 0.425. It showed a certain performance improvement over the LCV2 constructed with BLIP-VQA-Capfilt-large and Grounding DINO swin-T in the Answer Grounding task, where the latter obtained an IoU metric of 0.417. This is mainly attributed to the superior performance of Grounding DINO swin-B in open-world object localization tasks. The LCV2 with the VQA module based on the GittextVQA model exhibited the poorest performance in the experiment, with an IoU of 0.405. This is possibly due to the inferior visual question

answering performance of Git-large-VQAv2 compared to BLIP-VQA-Capfilt-large and the original Lens. The LCV2's VQA module, based on the BLIP-VQA-Capfilt-large model finetuned on the VizWiz Answer Grounding training set, achieved the best performance. For the BLIP-VQA-Large finetuning on the training set of VizWiz Answer Grounding, trainable bypassing was added to the Q and V matrices of its transformer framework, with the rank of the trained matrices being set to 8, the scaling factor to 16, and the dropout rate to 0.05, with a total of 150 epochs of training. The experimental results further indicate that finetuning contributes to enhancing the adaptability of the VQA module to the VizWiz Answer Grounding dataset. However, as these versions of the LCV2 provide visual answer grounding results in the form of Bbox, their performance is not particularly outstanding. Future efforts should aim to replace the OVD/REC module that offers finer grained grounding, enhancing the framework's answer grounding precisely for the visually impaired.

Table 7. A validation of the performance variation of the LCV2 on the VizWiz Answer Grounding dataset based on different VQA modules.

Models	IoU
LCV2 (lens-swinB)	0.424
LCV2 (Gittext-swinB)	0.405
LCV2 (BlipL-swinB)	0.425
LCV2 (BlipL-swinT)	0.417
LCV2 (FineTunedBlipL-swinB)	**0.430**

5. Analysis and Discussion

5.1. LCV2 Compared to Baseline Methods

The experiments on the GQA validation set demonstrated that the LCV2 significantly outperformed baseline approaches. As shown in the visual grounding results in Figure 7, the baseline methods ground their results by focusing on the attention paid by the model to the connections between textual semantics and visual content. The visual object boundary box is extracted by grounding the connected components of the attention map. This strategy effectively ensures the interpretability of the model's reasoning. However, it is susceptible to attention noise and lacks fine granularity, making it difficult to achieve the grounding precision of the LCV2's VG module, which operates as an expert model. This highlights an advantage of the LCV2: leveraging the advanced performance of domain-specific expert models to achieve superior results. The LCV2 not only utilizes advanced VQA models to predict the text answers to visual questions but also focuses on employing state-of-the-art OVD/REC models for the visual object grounding, thereby ensuring the accuracy of the visual question answering grounding.

The LCV2's performance on the CLEVR validation set was relatively poor. The reason is that the CLEVR dataset consists of computer-generated abstract geometric shapes and does not reflect real-world scenes. The LCV2's VQA module evaluated in this experiment was trained on real-world scenes and did not effectively generalize to the visual question answering tasks of the CLEVR dataset. Therefore, the evaluated LCV2 exhibited a poor generalization performance when dealing with such datasets. In Section 4.5 of the article, we finetuned the LCV2's VQA module on the CLEVR dataset using LoRA to further investigate the performance improvement of the LCV2 after finetuning the VQA module. Moreover, it is noteworthy that the LCV2 achieved the highest IoU and overlap recall rates on the CLEVR validation set, indicating fewer missed detections. However, its performance in terms of IoU and Overlap accuracy was rather ordinary, suggesting that the detection results might be accompanied by a higher false positive rate.

5.2. Impact of the VQA Module

In Section 4.5 of the experiments, we assessed the impact of different VQA modules on the experimental outcomes of LCV2. On the GQA validation set, the LCV2 equipped

with the Lens-based VQA module exhibited the poorest accuracy in the visual question answering. However, it should be noted that the responses produced by the Lens-based VQA module may not necessarily be inaccurate. Conversely, Lens demonstrated strong answer flexibility, often providing synonyms for the ground truth answer or alternative textual descriptions of the ground truth answer, as it leverages an LLM module to extract and infer the predictions from visual description modules. Moreover, Lens's predicted answers tended to be longer texts, as illustrated in Figure 9, which included more phrases about the visual content, leading the LCV2 to generate excessive bounding boxes not centered around the optimal areas. The models' computational complexity and parameter sizes are shown in Table 4. Lens has much higher computational costs and model parameters than the other models. This is due to Lens' modular design, which incorporates multiple expert models. BLIP-VQA-Capfilt-large, with its hybrid encoder–decoder architecture utilizing multiple modalities, achieved the best F1 score results for the VQA grounding with the least computational complexity and model parameters, suggesting that the LCV2 (BLIP) framework is more worthy of promotion and application.

The experiments in Section 4.5 on the CLEVR dataset evaluated the impact of fine-tuning the LCV2's VQA module on a domain-specific dataset. The knowledge from the VQA module was extended from real-world visual question answering to abstract 3D geometric shape scenes, demonstrating that the LCV2 can be finetuned to further adapt to specific scene applications. We assessed the stability of the method's inference process by analyzing the generalization performance of the LCV2 on different question types in the CLEVR validation set. As shown in Figure 10, the LCV2 demonstrated a relatively consistent performance across a range of question types. However, the method excels relatively in the attribute comparison questions, because the framework better captures the recognition and comparison of object attributes, exhibiting relatively strong capabilities in feature extraction and relational reasoning. The framework also achieves good F1 scores in numerical comparisons and querying attributes, showing strong capabilities in numerical logic understanding and attention to specific details. Additionally, the LCV2 is less precise in understanding the spatial relationships between the objects in complex scenes. In scenarios with numerous objects or subtle differences, it fails to effectively focus on relevant objects or features, resulting in a relatively mediocre performance on the count and existing type questions compared to other areas. The framework is better at recognizing attributes and basic numerical logic understanding than it is at fine visual perception and spatial relationship reasoning.

5.3. LCV2 on the VizWiz Answer Grounding

The LCV2 framework demonstrated a moderate performance on the VizWiz Answer Grounding dataset, with a significant influencing factor being the format of the grounding provided by our framework, which is in the form of bounding boxes (Bbox). In contrast, the ground truth labels provided by the VizWiz Answer Grounding dataset are in a more fine-grained image segmentation format, consisting of points along all edges of the answer grounding. This discrepancy leads to a negative impact on the experimental evaluation of the answer grounding task, even if our method provides accurate answer grounding information. However, this should not overshadow the advantages of the LCV2 approach, which, through its modular design, achieves pretraining-free and out-of-the-box functionality. Based on its plug-and-play module composition, it facilitates rapid deployment and iteration, allowing for easy customization for different tasks or requirements. For future improvements, we could substitute the OVD/REC module in our framework with a model that can offer grounding information in a segmentation format. This will enhance the performance on the VizWiz Answer Grounding dataset task.

6. Conclusions and Prospects

To enable a VQA grounding system in low computational resource settings, applicable to visual navigation for the visually impaired and broader human–machine interaction

scenarios, this paper proposes a pretraining-free universal framework. It is based on a frozen pretrained LLM to connect and transform the information between the VQA system and the OVD/REC system. In this setup, the VQA module predicts answers to visual questions, while the OVD/REC module provides feedback on visual cues grounding. The LLM transforms the original question and the text answer predicted by the VQA module into text content suitable for the OVD/REC model's grounding needs, enhancing the accuracy of the model in pinpointing visual details directed by the question. The modules of this framework can be easily adopted and replaced with the latest community models and methods without considering their pretraining processes and heterogeneous networks. The experiments on benchmark datasets, including GQA, CLEVR, and VizWiz answer grounding, validated the effectiveness of our proposed method.

In future work, the model framework's modules can be replaced with off-the-shelf models with larger parameters and a superior performance, assuming more powerful computational and storage resources are available. Consideration could also be given to more fine-grained visual grounding systems acting as the OVD/REC module to promote more precise grounding, such as those offering image segmentation. Furthermore, to further enhance the robustness of the large language model's output for the framework's task completion, future work could consider introducing supervisory prompt training [74] strategies, or memory mechanisms [75], etc., to the LLM. Additionally, the LCV2 could be extended to other intriguing domains, such as video and 3D-aware visual question answering grounding, generating valuable insights. For example, based on the Glance–Focus model [76], which mimics the human ability to quickly ground and understand using event memory, tasks can be extended to continuous frames or, by integrating reasoning with the PO3D-VQA model [77], the framework's three-dimensional scene perception capabilities could be achieved. The expansion would further facilitate the integration and coordination of the LLM and multimodal models, exploring intriguing directions for the LLM applications in the community.

7. Limitations

In this section, we summarize some limitations of the LCV2. Although the LCV2 avoids large-scale retraining through modular design, the introduction of a large language model for the intermediate text conversion steps increases the inference time, which is detrimental to the real-time performance.

Secondly, the LCV2 inherently processes visual and linguistic information in separate steps. Compared to end-to-end deep learning architectures, this segmented approach may sacrifice some depth and consistency in cross-modal understanding. LCV2 may not fully emulate the deep interaction present in end-to-end models, especially in tasks requiring deep semantic reasoning or complex visual relationship recognition.

Additionally, the LCV2 relies on a large language model to transform original question and answer texts into captions that are highly descriptive and contextually consistent. Although large language models possess strong capabilities in language generation, in certain complex or ambiguous question and answer scenarios, the model may not perfectly generate captions that are both precise and sufficiently focused on the relevant visual areas. The transformation process might involve information loss, semantic deviations, or oversimplification, which could affect the positioning accuracy of the subsequent OVD/REC models.

Author Contributions: Conceptualization, Y.C. and L.S.; methodology, Y.C.; software, Y.C.; validation, L.C. and Z.L.; formal analysis, Y.C. and L.S.; investigation, Z.L.; resources, Z.L. and L.C.; data curation, L.C.; writing—original draft preparation, Y.C.; writing—review and editing, L.S.; visualization, Y.C.; supervision, L.S.; project administration, L.S.; and funding acquisition, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: The project was supported by the Foundation for Science and Technology Program of State Grid—East China Branch (Grant NO. SGHD0000AZJS2310287).

Data Availability Statement: Publicly available datasets were analyzed in this study. The GQA datasets are available at https://cs.stanford.edu/people/dorarad/gqa/ (accessed on 6 January 2024). The CLEVR datasets can be found at https://cs.stanford.edu/people/jcjohns/clevr/ (accessed on 8 January 2024). The VizWiz-VQA-Grounding Dataset are sourced from https://vizwiz.org/tasks-and-datasets/answer-grounding-for-vqa/ (accessed on 12 January 2024). Additional data involved in this study are available on request from the corresponding authors.

Acknowledgments: We appreciate the support and financial assistance from the State Grid East China Branch for our work. We would also like to sincerely thank each and every one of the contributors of the open-source datasets that were used for this investigation.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Description and comparison of some methods related to VQA grounding.

Methods	Description	Advantages and Limitations
MAC-Caps [7]	Integrating a capsule network module with a query selection mechanism into the MAC VQA systems improved the model's answer grounding capabilities.	Pros: The introduced capsule network module enhances the attention and comprehension of the visual information mentioned in the question. Cons: The grounding of visual cues is provided by attention maps, which are subject to attentional interference and lack of fine granularity, affecting the accuracy of visual cue grounding.
DaVI [5]	The framework is a unified end-to-end system that employs a vision-based language encoder and a language-based vision decoder to generate answers and provide visual evidence based on visual question answering.	Pros: The approach integrates visual understanding and language generation, providing a unified solution that simultaneously addresses visual question answering and the visual evidence grounding of answers, enhancing the model's overall performance. Cons: The performance of the model depends on large-scale multimodal pretraining data, limiting its effectiveness in specific domains or on small datasets.
XMETER [60]	By integrating monolingual pretraining and adapter strategies, advanced English visual question answering models are extended to low-resource languages, generating corresponding bounding boxes for key phrases in multilingual questions.	Pros: The model effectively addresses the challenges of visual question answering tasks in low-resource language environments, demonstrating strong adaptability, efficiency, and performance. Cons: For questions involving multi-level logical reasoning (such as relational problems), the model lacks the robust semantic understanding required to fully grasp the task.
DDTN [10]	A dual-decoder transformer network is proposed, which efficiently predicts language answers and corresponding visual instance grounding in visual question answering by integrating region and grid features and employing an instance query-guided feature processing strategy.	Pros: The model employs a unique dual-decoder design, which facilitates the separate handling of language comprehension and visual grounding tasks. Cons: The model has a limited ability to precisely ground and segment objects with complex contours and complex backgrounds.
P2G [61]	The reasoning process is enhanced with multimodal cues by utilizing external expert agents to perform the real-time grounding of key visual and textual objects.	Pros: Based on agent queries of visual or textual cues, the P2G model can perform more purposeful reasoning. Cons: For extremely complex or atypical scenarios, such as densely stacked text or highly abstract visual elements, the model still struggles with accurate understanding and reasoning.
LCV2	The framework utilizes a LLM to connect the VQA and VG modules, based on a modular design, to circumvent the challenges posed by extensive pretraining in modeling.	Pros**:** Leveraging the generalizable knowledge of expert models allows for out-of-the-box functionality without the need for any further training in modeling**.** Cons**:** The non-end-to-end design may compromise the depth and consistency of the cross-modal understanding to some extent.

Table A2. Comparison of reasoning performance across various tasks for several representative large language models.

Model	Params	Pretrained Data Scale	MMLU	MATH	GSM8K	BBH	CEval
GPT-3.5 [47]	175B	-	70.0	-	57.1	-	54.4
GPT-4 [48]	-	-	86.4	42.5	92.0	-	68.7
ChatGLM [53]	6.2B	1T tokens	36.9	-	4.82	-	38.9
OPT [68]	175B	180B tokens	25.2	-	-	-	25.0
PALM [52]	8B/62B/540B	780B tokens	62.9	8.8	56.9	62.0	-
Flan-T5 [45]	0.75B/3B/11B	780B tokens	48.6	-	16.1	41.4	-

References

1. Lu, S.; Liu, M.; Yin, L.; Yin, Z.; Liu, X.; Zheng, W. The multi-modal fusion in visual question answering: A review of attention mechanisms. *PeerJ Comput. Sci.* **2023**, *9*, e1400. [CrossRef] [PubMed]
2. Chen, C.; Anjum, S.; Gurari, D. Grounding answers for visual questions asked by visually impaired people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 19098–19107.
3. Massiceti, D.; Anjum, S.; Gurari, D. VizWiz grand challenge workshop at CVPR 2022. *ACM SIGACCESS Access. Comput.* **2022**, 1. [CrossRef]
4. Zeng, X.; Wang, Y.; Chiu, T.-Y.; Bhattacharya, N.; Gurari, D. Vision skills needed to answer visual questions. *Proc. ACM Hum. Comput. Interact.* **2020**, *4*, 149. [CrossRef]
5. Liu, Y.; Pan, J.; Wang, Q.; Chen, G.; Nie, W.; Zhang, Y.; Gao, Q.; Hu, Q.; Zhu, P. Weakly-Supervised Grounding for VQA with Dual Visual-Linguistic Interaction. In Proceedings of the CAAI International Conference on Artificial Intelligence, Fuzhou, China, 22–23 July 2023; pp. 156–169.
6. Xiao, J.; Yao, A.; Li, Y.; Chua, T.S. Can I trust your answer? visually grounded video question answering. *arXiv* **2023**, arXiv:2309.01327.
7. Urooj, A.; Kuehne, H.; Duarte, K.; Gan, C.; Lobo, N.; Shah, M. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, Nashville, TN, USA, 19–25 June 2021; pp. 8465–8474.
8. Khan, A.U.; Kuehne, H.; Gan, C.; Lobo, N.D.V.; Shah, M. Weakly supervised grounding for VQA in vision-language transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 652–670.
9. Le, T.M.; Le, V.; Gupta, S.; Venkatesh, S.; Tran, T. Guiding visual question answering with attention priors. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 4381–4390.
10. Zhu, L.; Peng, L.; Zhou, W.; Yang, J. Dual-decoder transformer network for answer grounding in visual question answering. *Pattern Recogn. Lett.* **2023**, *171*, 53–60. [CrossRef]
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D* **2020**, *404*, 132306. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Malinowski, M.; Fritz, M. A multi-world approach to question answering about real-world scenes based on uncertain input. *arXiv* **2014**, arXiv:1410.0210.
16. Ren, M.; Kiros, R.; Zemel, R. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Adv. Neural Inf. Process. Syst.* **2015**, *1*, 5.
17. Yu, Z.; Yu, J.; Fan, J.; Tao, D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1821–1830.
18. Ben-Younes, H.; Cadene, R.; Cord, M.; Thome, N. MUTAN: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2612–2620.
19. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
20. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, arXiv:1908.03557.

21. Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; Wang, L. Git: A generative image-to-text transformer for vision and language. *arXiv* **2022**, arXiv:2205.14100.
22. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
23. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
24. Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
25. Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; Sun, T. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv* **2023**, arXiv:2306.17107.
26. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
27. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.-F. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, Nashville, TN, USA, 19–25 June 2021; pp. 14393–14402.
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
29. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10965–10975.
30. Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; Xu, H. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9125–9138.
31. Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; Berg, T.L. Mattnet: Modular attention network for referring expression comprehension. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1307–1315.
32. Hong, R.; Liu, D.; Mo, X.; He, X.; Zhang, H. Learning to compose and reason with language tree structures for visual grounding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *44*, 684–696. [CrossRef]
33. Shi, F.; Gao, R.; Huang, W.; Wang, L. Dynamic MDETR: A dynamic multimodal transformer decoder for visual grounding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 1181–1198. [CrossRef] [PubMed]
34. Yang, Z.; Chen, T.; Wang, L.; Luo, J. Improving one-stage visual grounding by recursive sub-query construction. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIV 16. pp. 387–404.
35. Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; Ji, R. Seqtr: A simple yet universal network for visual grounding. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 598–615.
36. Subramanian, S.; Merrill, W.; Darrell, T.; Gardner, M.; Singh, S.; Rohrbach, A. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv* **2022**, arXiv:2204.05991.
37. He, R.; Cascante-Bonilla, P.; Yang, Z.; Berg, A.C.; Ordonez, V. Improved Visual Grounding through Self-Consistent Explanations. *arXiv* **2023**, arXiv:2312.04554.
38. Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; Liu, J. Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6616–6628.
39. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the European Conference on Computer Vision, Virtual, Glasgow, UK, 23–28 August 2020; pp. 104–120.
40. Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; Lu, H. Universal instance perception as object discovery and retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 15325–15336.
41. Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R.K.; Mahadevan, V.; Manmatha, R. Polyformer: Referring image segmentation as sequential polygon generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18653–18663.
42. Xuan, S.; Guo, Q.; Yang, M.; Zhang, S. Pink: Unveiling the power of referential comprehension for multi-modal llms. *arXiv* **2023**, arXiv:2310.00582.
43. Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; Kembhavi, A. Unified-io: A unified model for vision, language, and multi-modal tasks. In Proceedings of the Eleventh International Conference on Learning Representations, Virtual, 25–29 April 2022.
44. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 23318–23340.

45. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S. Scaling instruction-finetuned language models. *arXiv* **2022**, arXiv:2210.11416.
46. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
47. Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv* **2023**, arXiv:2303.10420.
48. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
49. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
50. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv* **2021**, arXiv:2107.02137.
51. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F. Qwen technical report. *arXiv* **2023**, arXiv:2309.16609.
52. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **2023**, *24*, 1–113.
53. Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X. Glm-130b: An open bilingual pre-trained model. *arXiv* **2022**, arXiv:2210.02414.
54. Hudson, D.A.; Manning, C.D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6700–6709.
55. Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2901–2910.
56. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
57. Chen, C.; Anjum, S.; Gurari, D. VQA Therapy: Exploring Answer Differences by Visually Grounding Answers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 15315–15325.
58. Hudson, D.A.; Manning, C.D. Compositional attention networks for machine reasoning. *arXiv* **2018**, arXiv:1803.03067.
59. Pan, J.; Chen, G.; Liu, Y.; Wang, J.; Bian, C.; Zhu, P.; Zhang, Z. Tell me the evidence? Dual visual-linguistic interaction for answer grounding. *arXiv* **2022**, arXiv:2207.05703.
60. Wang, Y.; Pfeiffer, J.; Carion, N.; LeCun, Y.; Kamath, A. Adapting Grounded Visual Question Answering Models to Low Resource Languages. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2595–2604.
61. Chen, J.; Liu, Y.; Li, D.; An, X.; Feng, Z.; Zhao, Y.; Xie, Y. Plug-and-Play Grounding of Reasoning in Multimodal Large Language Models. *arXiv* **2024**, arXiv:2403.19322.
62. Dou, Z.-Y.; Kamath, A.; Gan, Z.; Zhang, P.; Wang, J.; Li, L.; Liu, Z.; Liu, C.; LeCun, Y.; Peng, N. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32942–32956.
63. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv* **2023**, arXiv:2303.05499.
64. Xie, C.; Zhang, Z.; Wu, Y.; Zhu, F.; Zhao, R.; Liang, S. Described Object Detection: Liberating Object Detection with Flexible Expressions. *arXiv* **2024**, arXiv:2307.12813.
65. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 16 January 2024).
66. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
67. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
68. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.V. Opt: Open pre-trained transformer language models. *arXiv* **2022**, arXiv:2205.01068.
69. Berrios, W.; Mittal, G.; Thrush, T.; Kiela, D.; Singh, A. Towards language models that can see: Computer vision through the lens of natural language. *arXiv* **2023**, arXiv:2306.16410.
70. GQA: Visual Reasoning in the Real World—Stanford University. Available online: https://cs.stanford.edu/people/dorarad/gqa/download.html (accessed on 6 January 2024).
71. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
72. Answer Grounding for VQA—VizWiz. Available online: https://vizwiz.org/tasks-and-datasets/answer-grounding-for-vqa/ (accessed on 12 January 2024).

73. Hu, R.; Andreas, J.; Darrell, T.; Saenko, K. Explainable neural computation via stack neural module networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 53–69.
74. Billa, J.G.; Oh, M.; Du, L. Supervisory Prompt Training. *arXiv* **2024**, arXiv:2403.18051 2024.
75. De Zarzà, I.; de Curtò, J.; Calafate, C.T. Socratic video understanding on unmanned aerial vehicles. *Procedia Comput. Sci.* **2023**, *225*, 144–154. [CrossRef]
76. Bai, Z.; Wang, R.; Chen, X. Glance and Focus: Memory Prompting for Multi-Event Video Question Answering. *arXiv* **2024**, arXiv:2401.01529.
77. Wang, X.; Ma, W.; Li, Z.; Kortylewski, A.; Yuille, A.L. 3D-Aware Visual Question Answering about Parts, Poses and Occlusions. *arXiv* **2024**, arXiv:2310.17914.