Acknowledging Focus Ambiguity in Visual Questions

Chongyan Chen[1†], Yu-Yun Tseng[2†], Zhuoheng Li[2], Anush Venkatesh[2], Danna Gurari[1,2]
[1] University of Texas at Austin [2] University of Colorado Boulder
[†] denotes equal contribution (and so shared first authorship)

Abstract

No published work on visual question answering (VQA) accounts for ambiguity regarding where the content described in the question is located in the image. To fill this gap, we introduce VQ-FocusAmbiguity, the first VQA dataset that visually grounds each plausible image region a question could refer to when arriving at valid answers. We next analyze and compare our dataset to existing datasets to reveal its unique properties. Finally, we benchmark modern models for two novel tasks related to acknowledging focus ambiguity: recognizing whether a visual question has focus ambiguity and locating all plausible focus regions within the image. Results show that the dataset is challenging for modern models. To facilitate future progress on these tasks, we publicly share the dataset with an evaluation server at https://vizwiz.org/tasks-and-datasets/focus-ambiguity-in-visual-questions/.

1. Introduction

Ambiguous language is a common part of communication. It entails using vague words or phrases that can be interpreted in multiple plausible ways, ideally alongside context clarifying the intended meaning. For example, when a three-year-old asks "What is this?", we can understand the meaning if the child simultaneously points to an item (e.g., a red pomegranate). However, context is not always provided to clarify the intended meaning of a question, as exemplified in **Figure 1**. In such cases, *the language in the question could refer to multiple plausible visual regions*. We call this *focus ambiguity in visual questions*, or more concisely "ambiguous questions" and "question ambiguity".

Our paper is motivated by the belief that a VQA system *should* notify users when there is question ambiguity and then facilitate them to arrive at the desired interpretation. The possible repercussions from VQA services not providing such information can be grave, potentially inflicting adverse social, professional, legal, financial, and personal consequences. For instance, imagine if the answer to the question in **Figure 1** about "What is the cleaning product?" leads a blind person to use window cleaner to wash their dishes they will use for eating instead of the dish soap. Alternatively, imagine replacing the question in **Figure 1** with "What is the medicine?" when instead three pill bottles are visible. The current obstacle for developing ambiguity-aware VQA solutions is that no benchmark dataset exists for establishing to what extent VQA models are aware of question ambiguity and so where further improvement is needed.

To fill this gap, we introduce the first dataset goal-oriented towards focus ambiguity in visual questions. Called VQ-FocusAmbiguity, it consists of 5,500 examples and segments (aka, grounds) all plausible image regions to which the language in each question could refer. The dataset has a nearly even distribution between ambiguous and unambiguous examples. We also characterize this

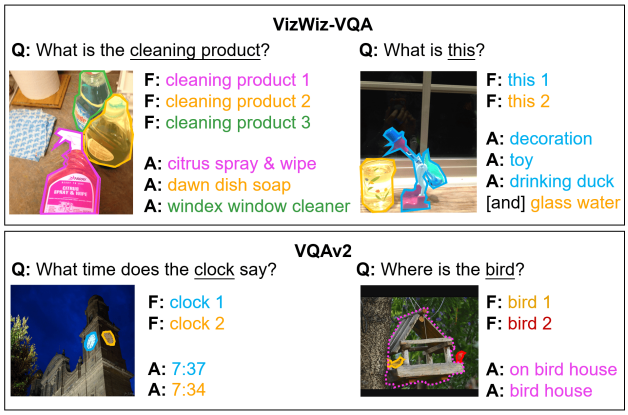

Figure 1. Visual questions with focus ambiguity, meaning language in the questions (underlined words/phrases) refer to multiple plausible image regions (aka, segmentations). 'Q' denotes the question, 'F' denotes the focus specified in the question, and 'A' denotes the answer. We created a new dataset that originates from four sources to represent ambiguous and unambiguous questions, with two of the sources originating from existing VQA datasets [15, 22] and exemplified in this figure. As shown, the location of the question grounding and answer grounding can match (in which case, matching colors are used for all segmentations and their associated text) and can differ (bottom right example, where an additional answer segmentation is shown with a dashed line).

dataset and how it relates to a related *answer* grounding dataset. Crucially, as exemplified in **Figure 1**, our analysis underscores the importance of disentangling questions as a source of ambiguity, since the focus of language in *questions* can differ from visual evidence showing *answers*. Additional scenarios where the question and answer groundings can differ include for questions about relations (e.g., "What is above the mirror?", with the question grounding of the mirror, and answer grounding of objects above the mirror) and activities (e.g., "What is the person doing?", with the question grounding of a person and the answer grounding of a frisbee). Finally, we show that modern models perform poorly for the tasks of (1) recognizing whether a visual question has focus ambiguity and (2) locating all plausible image regions to which the content described in the question could refer. Our fine-grained analysis reveals where models struggle, and so provides valuable insights for future efforts in model development.

Success in developing ambiguity-aware solutions can immediately benefit today's users of VQA services, spanning blind and sighted individuals, who already regularly ask visual questions using *mobile phone apps* (e.g., Be My AI, Microsoft's Seeing AI), *smart glasses* (e.g., Meta's Ray Bans, Envision AI), and the *web* (e.g., Q&A platforms such as Stack Exchange). It would enable AI agents to alert users when there is question ambiguity as well as interactively guide users towards disambiguating their intent by having them specify which from all plausible image regions the question is referring to. This work can also support enhanced reasoning abilities of AI agents, by encouraging an intermediate step of determining a question's focus in an image towards deciding what answer to predict. Finally, we expect this work will serve as a pioneering example for addressing focus ambiguity for *other vision-language tasks* (e.g., image captioning [11], visual storytelling [34], language-guided image editing [37]), and *across more modalities* (e.g., focus ambiguity for questions asking about visual, audible, tactile, or olfactory content).

2. Related Work

Automatically Acknowledging Ambiguity in Question Answering. Automated systems already account for several types of ambiguity in question answering. Most focus on a purely language setting [17, 27, 38, 49], such as when a word has multiple plausible meanings (e.g., "revolting" can mean rebelling or disgusting). A few focus on multimodal VQA [4, 19, 40, 46, 50, 53], by accounting for *when and why* different *answers* are observed from different people (e.g., due to subjectivity and differing levels of granularity). Our work complements prior work by being the **first to disentangle the *questions* themselves as a source of ambiguity**, thereby helping clarify *how* different answers can arise by showing the reasoning process from a question that leads to the answers.

VQA and VQA Grounding Datasets. Many datasets exist for VQA [2, 7, 15, 22, 36, 45] and locating the visual evidence showing where *answers* to visual questions reside [5, 6, 14, 24, 29, 51, 55]. One dataset, Visual7W [55], even locates where the language in *questions* refers to in images. Extending prior work, we introduce the **first dataset goal-oriented to focus ambiguity, with examples showing when language in *questions* refers to *multiple* plausible image regions**.

Natural Language Localization. More generally, this work contributes to existing literature on locating linguistic expressions in an image, which is already explored for tasks like object detection [56], instance segmentation [18, 23], referring expression comprehension [8, 41], and described object detection [44, 52]. Complementing these tasks, our work **locates where language in a question refers to in an image, which yields diverse linguistic expressions**, including vague terms such as "this", "it", and "her" (e.g., third person pronouns, singular demonstrative pronouns).

3. VQ-FocusAmbiguity Dataset

3.1. Dataset Creation

Each example in our dataset has three parts: an image, a question, and segmentations for all regions that could be the focus of the question. We created the dataset by extending four diverse sources, that are described in **Table 1**. The visual data spans content that (1) shows a single object and complex scenes; (2) comes from sighted and visually impaired photographers, and (3) has objects at various locations and of many sizes. Questions ask about many subjects—including about objects and their parts—as well as about their relationships and actions, using vague terms (e.g., "this"), specific categories (e.g., "bus"), and detailed descriptions (e.g., "person in blue next to the car").

3.1.1. Extensions of Segmentation Datasets

Most examples are derived from two entity segmentation datasets, which already provide images with segmentations.

Data Source. We leverage test sets of PACO [42] and MSRA-B [33], which both contain images scraped from online image-sharing platforms. PACO uses the COCO images [31], which each show a complex scene with two or more objects. PACO provides exhaustive segmentations for all instances belonging to 75 object and 200 part categories. MSRA-B, in contrast, is a salient object segmentation dataset designed to contain a single foreground object per image, agnostic to the object category [26].

Data Filtering. For PACO, we randomly sampled 2,272 examples. For MSRA-B, we focused on the 626 images

Dataset	Image Source	Question Source	Ambiguity Labels	Segmentations	% Unambig (#)
PACO [42]	COCO (2017) [18, 31]	**Synthesized + Workers**	**Workers**	**Workers***	50% (2,272)
MSRA-B [33]	MSRA-B [33]	**Synthesized**	STATIC [20]	Workers	100% (626)
AnswerTherapy [6]	COCO (2015) [31]	Workers [31]	**Workers**	**Workers**	47% (82)
	VizWiz [22]	Visually Impaired People [22]			53% (83)

Table 1. Description of the four data sources used in VQ-FocusAmbiguity. Entries in bold represent new annotations created by our team. (* denotes when annotators selected between candidate segmentations rather than creating them from scratch; Unambig = Unambiguous).

that both lacked human faces and adult content (we determined this using the GPT-4o model [39]) and were flagged as containing "a single, noncontroversial foreground object of interest" [21].

Data Annotation. We next established questions with language referring to the segmentations.

For PACO, we achieved this using a home-grown annotation interface that showed the image with all available segmentations overlaid on the image. The interface first prompted the annotator to generate a question by presenting a list of AI-suggested questions and letting the annotator choose between either (1) authoring a question from scratch, (2) selecting a suggested candidate question as is, or (3) selecting a suggested candidate question after refining it.[1] Next, the user was prompted to select all segmentations to which the question could be grounded.

For MSRA-B, we only generated unambiguous questions and used the single segmentation per image as the question's focus. We automatically generated the questions, using variants of the most common question asked by people with vision impairments [22]: "What is this?"

3.1.2. Extensions of Visual Question Answering Datasets

The remaining examples extend two VQA datasets.

Data Source. We extend VizWiz-VQA [22] and VQAv2 [16]. VizWiz-VQA represents an authentic use case where people with visual impairments asked questions about images they took. VQAv2 is the most popular VQA dataset for model benchmarking and was created by asking people to make up questions about images that would "stump a robot". We focused on the 4,440 examples from these sources contained in the publicly-available splits of the AnswerTherapy dataset [6] (i.e., its training and validation sets) to enable comparison between its answer groundings and our question groundings.

Data Annotation. We established a trusted annotation protocol through three iterative refinement steps, discussed in the supplementary materials. Following this protocol, in-house annotators labeled every example from the data source by (1) locating the phrase in the question that needed to be grounded to the image in order to answer the question and (2) indicating whether there was ambiguity where the phrase is referring to in the image. This culminated in 165 ambiguous questions and 4,275 unambiguous questions, with 85% (i.e., 3792/4440) belonging to visual questions with a single answer grounding and 15% (i.e., 648/4440) belonging to visual questions with multiple answer groundings. The annotators then segmented all 165 ambiguous questions and 165 randomly sampled unambiguous questions, locating all regions the relevant phrase could focus on in the images.

Data annotation culminated in 5,500 visual questions with 12,880 instance segmentations and 5,500 classification labels (the binary flag for ambiguity is inferred from the number of segmentations).[2] Examples are nearly evenly distributed between containing and lacking question ambiguity, with 2,437 and 3,063 examples respectively.

3.1.3. Dataset Splits

Consistent with recent published VQA datasets [7, 25, 30, 35, 47, 54], we split this dataset to support zero/few-shot learning settings. This recent trend reflects that state-of-the-art performance regularly arises from foundation models in such settings; e.g., Frozen in 2021 [48], Flamingo in 2022 [1], ViTiS in 2023 [13], and LLaVA-v1.6 in 2024 [32]. For both the training and validation sets, we randomly sampled 10 unambiguous and 10 ambiguous examples from each data source to increase domain diversity. With four sources for unambiguous questions (PACO, MSRA-B, AnswerTherapy-VQAv2, and AnswerTherapy-VizWiz) and three sources for ambiguous questions (PACO, AnswerTherapy-VQAv2, and AnswerTherapy-VizWiz), we end up with 70, 70, and 5,360 examples in our training, validation, and test splits respectively.

3.2. Dataset Analysis

Questions. We first characterize how questions compare for examples with versus without focus ambiguity. We provide analysis with respect to the entire VQ-FocusAmbiguity dataset as well as with respect to each data source.

Statistics regarding how many words are in the questions are shown in **Figure 2(a)**. Overall, we observe a tendency

[1] When generating ambiguous questions, annotators chose options 1, 2, and 3 for 55%, 31%, and 14% respectively. For unambiguous questions, annotators chose options 1, 2, and 3 for 43%, 43%, and 14% respectively.

[2] To facilitate future research, we also publicly-share the metadata about the additional 4,110 examples from the VQA datasets that we flagged as unambiguous but did not segment.

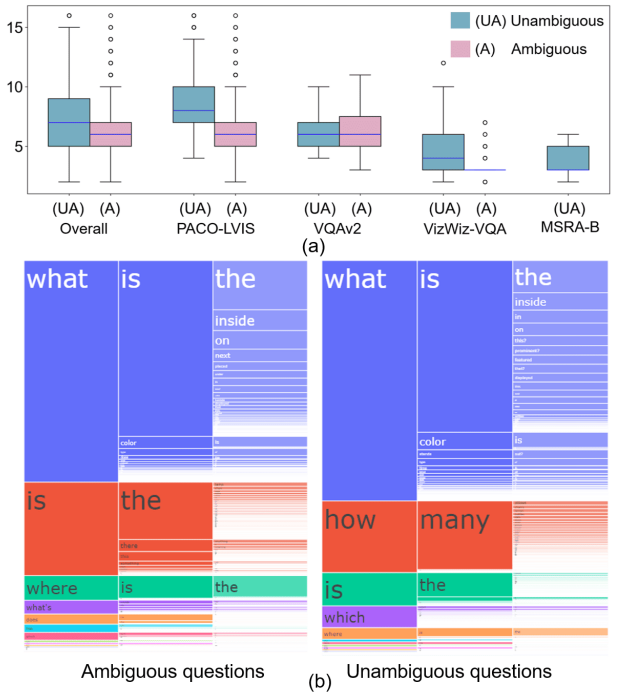

Figure 2. Analysis of questions in the dataset. (a) Box plot showing the number of words in questions that are unambiguous (UA) versus ambiguous (A), overall and for each data source (outliers are omitted for improved readability). (b) Icicle chart showing the first three words for all questions with and without question ambiguity. Each rectangle size is proportional to the number of questions with that word/phrase, with the left column showing the first word and each subsequent column showing a subsequent word.

for unambiguous questions to contain more words, particularly for examples from PACO. We hypothesize this correlation stems from extra words providing additional context that disambiguates the intended questions. Unambiguous questions also exhibit greater variation in length, as reflected by a higher standard deviation of 3.13 for unambiguous questions versus 1.97 for ambiguous questions.

We next characterize common linguistic patterns in questions by visualizing the distribution of their first three words in **Figure 2**(b). The key distinction between both question types is that questions *with* focus ambiguity more often begin with "Is the" while questions without focus ambiguity more often start with "How many". Intuitively, it makes sense that ambiguity is more likely to arise when a question is framed in a singular form ("Is a") rather than a plural form ("How many"), since the former does not permit acknowledging multiple image regions (e.g., "What color is the kite?" versus "What color are the kites?"). To further investigate this intuition, we utilized NLTK to flag whether any word in each question contains a plural noun (i.e., a plural common noun or plural proper noun). Supporting our hypothesis, we found that *unambiguous questions* are

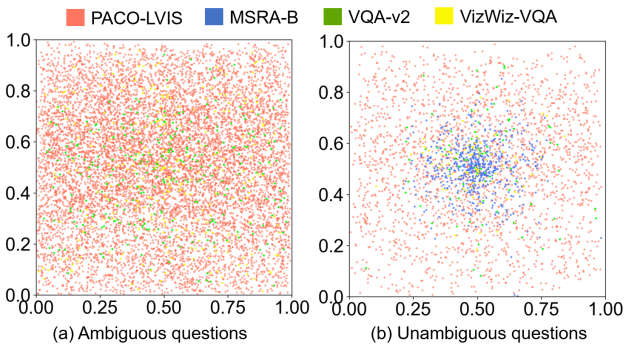

Figure 3. Location of each instance segmentation using normalized center of mass coordinates (x, y), for (a) all ambiguous questions and (b) all unambiguous questions. While both types of questions have instance segmentations located at a diversity of positions, unambiguous questions are biased to the center of images.

more than three times as likely to contain a plural noun than *ambiguous questions*, with 23.8% versus 4.7% respectively. This trend was more pronounced for PACO and for VQAv2, where questions were typed, than for VizWiz-VQA where the questions were initially spoken and subsequently transcribed.[3] Despite this slight difference in trends, the presence of plural nouns is not sufficient alone to determine whether there is question ambiguity.

Segmentations. We next characterize how segmentations compare for those with versus without focus ambiguity, again providing analysis for all of VQ-FocusAmbiguity and with respect to each data source.

We report the position of instance segmentations by computing the center of mass of each instance segmentation with respect to the entire image. Each coordinate can range from 0 to 1, and we normalize all images to ensure (x,y) values are comparable to each other. Results are shown in **Figure 3**. There is a greater bias for instance segmentations lacking focus ambiguity to be located in the center of an image, likely due to the salient object data source MSRA-B. However, we observe that unambiguous questions can also manifest the diverse locations typical for ambiguous questions, particularly for the unambiguous examples coming from the more complex images in PACO.

Summative statistics regarding how many instance segmentations are associated with each question are shown in **Figure 4(a)**. We exclude unambiguous examples since, by definition, they contain one segmentation. We observe similar statistics across all data sources, with an overall median of 3 and mean of 4 segmentations per ambiguous question.

We next measure the fraction of image pixels occupied by each instance segmentation. Results are shown in **Fig-**

[3]For PACO, plural nouns were found for 4.7% of ambiguous questions and 30.9% for unambiguous questions versus 4.2% and 17.6% for VQA-Therapy (7.7% and 26.8% for VQA-v2; 0% and 8% for VizWiz-VQA).

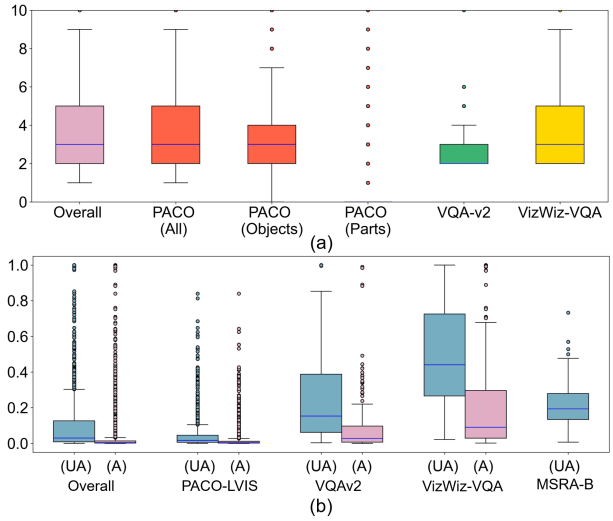

Figure 4. Box plots characterizing (a) the number of instance segmentations associated with each ambiguous example, with outliers omitted for improved readability since PACO can exceed 30 instance segmentations and (b) the fraction of pixels occupied by each instance segmentation for all examples.

ure 4(b). Unambiguous questions tend to have segmentations occupying a larger portion of the image, likely because the tendency in such cases is for images to feature a single, dominant salient object. However, we do observe unambiguous examples with very small coverage like that observed for ambiguous questions, particularly for the PACO data source. Consequently, properties of an instance's image coverage alone is insufficient for predicting whether there is question ambiguity.

Additionally, we analyze the prevalence of objects versus parts for serving as the instance segmentations for the PACO data source. Within PACO, 81.4% of instance segmentations are of only objects, 15.8% are of only parts, and 2.8% feature a mix of objects and parts.

Question Groundings Versus Answer Groundings. We flagged for all examples from the AnswerTherapy dataset (i.e., 330) whether the instance segmentations in our *question* groundings match the *answer* groundings. We observed different trends for the different types of questions. For the ambiguous questions, 79% (131 out of 165) had groundings that are *different* for the question and answers. In contrast, 64% (106 out of 165) of unambiguous questions had groundings that are *matching* for the questions and answers. Examples of both scenarios are shown in **Figure 5**. These findings reinforce the importance of locating a question's focus as an important, independent stepping stone towards providing users of VQA services all valid answers.

Reasons for Focus Ambiguity. For 265 examples with question ambiguity, we manually coded the reasons for question ambiguity. We used all 91 examples from VQAv2,

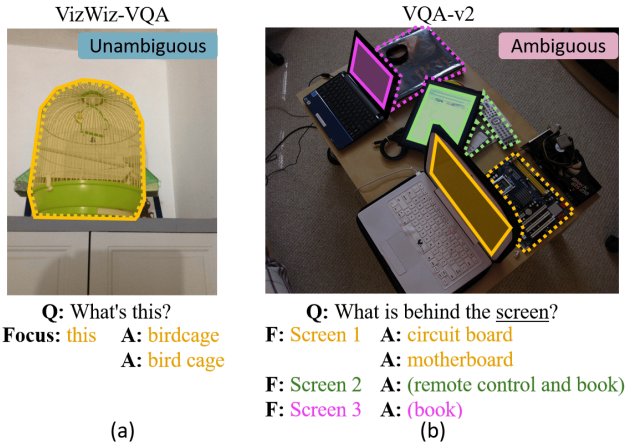

Figure 5. Examples of groundings for the question and answers that (a) match and (b) differ. The answers in parentheses are not provided in the AnswerTherapy dataset.

74 from VizWiz, and a random sample of 100 from PACO. We identified two primary reasons:

- **Multiple instances of the same category** account for 61.5% overall, with 84.6% (i.e., 77) in VQAv2, 4% (i.e., 3) in VizWiz-VQA, and 83% (i.e., 83) in PACO. An example is "What is next to the mirror?" when multiple mirrors are present.
- **Multiple instances of different categories** account for 31% overall, with 0.1% (i.e., 9) in VQAv2, 89% (i.e., 66) in VizWiz-VQA, and 9% (i.e., 9) in PACO. These usually happen when the questions are vague. Examples include "What is this?" and "What is outside the window?"

Other rare reasons include (a) perspective ambiguity (e.g., "Who is pulling the other side?"), (b) subjectivity (e.g., "What is the most distinctive feature on the building?"), (c) language ambiguity (e.g., "What is in the picture?" while it can either refer to the entire image or a painting in the image), and (d) multiple plausible parts for the same entity (e.g., "What is the part of the lamp that is fully visible?").

4. Model Benchmarking

We now benchmark models on VQ-FocusAmbiguity for two novel tasks: (1) recognizing whether a visual question has focus ambiguity and (2) locating all image regions that could be the question's focus.

4.1. Recognizing Questions with Focus Ambiguity

Models. We evaluate four foundation models. Three are top-performers on Arena-Vision [10] and MMMU benchmarks [54]: GPT-4o [39], InternVL2-Llama3-76B [9], and Qwen2.5-VL-7B-Instruct [3]. The fourth is Molmo-7B [12], a state-of-the-art language grounding model.

Prompts. We used five prompts for each model, resulting in 20 tested model variants. Three prompts involved no

supervision (i.e., zero-shot) while two incorporated a small number of examples to ideally boost performance (i.e., in-context few-shot learning). They are as follows:

- *Zero-shot (ZS)*: combines the definition of question ambiguity and the task objective.
- *Zero-shot chain of thought (ZS-CoT)*: facilitates the model's reasoning by augmenting ZS with the instruction "please think step by step".
- *Zero-shot enhanced chain of thought (ZS-ECoT)*: facilitates the model's reasoning by augmenting ZS with structured guidance (i.e., four steps) on how to perform our novel task (i.e., prompt decomposition).
- *Few-shot (FS)*: augments ZS with an ambiguous and an unambiguous example. A textual description is used for each example image to maintain evaluation consistency, as some models don't support multi-image input.
- *Few-shot enhanced chain of thought (FS-ECoT)*: augments FS with prompt decomposition from ZS-ECoT.

All prompts specify *please only answer "ambiguous" or "unambiguous"*, guiding the models to generate a one-word answer for binary classification. However, since generative models can and did produce arbitrary text beyond what was requested, we applied post-processing to categorize all outputs into three possible categories: "ambiguous", "unambiguous", and "undecided".

Evaluation Metrics. We employ four metrics. Two are standard binary classification metrics: accuracy and weighted f1 score. The third is "positive rate", measuring the percentage of positive predictions (i.e., predicting there *is* question ambiguity) to reveal potential biases in model predictions. The last metric is "undecided rate", which is the fraction of all examples with "undecided" predictions.

Overall Performance. Results are shown in **Table 2**. Overall, all models perform poorly, especially with respect to accuracy and F1 scores. This underscores that our dataset offers a challenging problem for the research community.

Our results also offer insights into strategies that can boost performance. First, we observe that facilitating models' reasoning abilities through chain-of-thought prompts (i.e., CoT and ECoT methods) leads to considerable performance gains across all but one model (i.e., InternVL2). For instance, ZS-CoT prompting boosts Molmo-7B's accuracy by 18.4 percentage points (pp), and ZS-ECoT yields an 8.3 pp improvement for Qwen2.5-VL. These gains elevate the performance of these 7B models to match or even surpass that of the much larger InternVL2 model (76B). We hypothesize the disparity observed for InternVL2 stems from the pretraining data, where both Qwen2.5-VL and Molmo were trained on region-level counting and pointing tasks included in the PixMo dataset [12], while InternVL2 was not trained. Such tasks are inherently related to ambiguity recognition: counting a single focus regions corresponds to an unambiguous question, and counting multiple focus regions indicate ambiguity. The positive rate scores provides evidence supporting our hypothesis, as InternVL2 consistently favors negative (i.e., unambiguous) predictions across prompting strategies while Molmo-7B and Qwen2.5-VL tend to classify questions as ambiguous under zero-shot conditions while adopting a more balanced perspective when guided by reasoning-based prompts. Together, these findings underscore the **complementary importance of both the prompting strategy and training data in enhancing vision-language models' ability to recognize question ambiguity**.

Model	Prompt	Acc.	F1	Pos.	UR
GPT-4o (over 200B)	ZS	67.7	67.4	63.5	1.9
	ZS-CoT	**69.6**	**69.8**	53.3	3.0
	ZS-ECoT	68.4	68.6	46.8	2.8
	FS	60.0	58.0	74.0	0.6
	FS-ECoT	64.9	65.1	45.4	2.8
InternVL2 (76B)	ZS	55.0	53.1	28.5	2.7
	ZS-CoT	**56.7**	**54.8**	27.9	4.1
	ZS-ECoT	54.9	53.2	29.6	3.1
	FS	52.4	51.9	36.8	1.1
	FS-ECoT	53.3	51.8	31.1	1.4
Qwen2.5-VL (7B)	ZS	57.2	53.5	79.1	0.0
	ZS-CoT	63.8	62.9	67.1	0.1
	ZS-ECoT	**65.5**	**65.3**	59.0	0.0
	FS	53.6	46.1	88.3	0.0
	FS-ECoT	59.0	55.8	78.1	0.0
Molmo (7B)	ZS	38.5	21.9	99.5	0.1
	ZS-CoT	**56.9**	**57.1**	48.1	6.9
	ZS-ECoT	45.7	42.6	75.0	1.0
	FS	41.8	24.6	100.0	0.0
	FS-ECoT	49.5	48.9	64.2	3.4

Table 2. Performance of 20 model variants for question ambiguity recognition with respect to accuracy (Acc.), weighted f1 score (F1), positive rate (Pos.), and undecided rate (UR) as percentages.

Fine-Grained Analysis. We next analyze the performance of the models with respect to data source and answer length. For data sources, we consider all image sources in VQA-FocusAmbiguity. For answer length, we categorize answers as short or long, where short answers contain one word (ideally "ambiguous" or "unambiguous", as instructed in the prompt) and all other answers are long. Results are shown in **Figure 6**.

With respect to *data source*, we observe models exhibit similar performance on the three balanced datasets (i.e., PACO, VQA-v2, VizWiz-VQA) and different performance on the highly imbalanced MSRA-B dataset (i.e., only contains unambiguous examples). For instance, across all five tested prompts, InternVL2 consistently performs *best* on MSRA-B while GPT-4o and Qwen consistently performs

worst on MSRA-B in comparison to the other three sources. Altogether, these findings underscore the models' resilience to variations in image and question sources, while also highlighting they bring different prediction biases.

With respect to *answer length*, while we generally observe similar performance for short and long answers, a notable exception is for the best-performing model where performance for longer answers surpasses that of short answers when using chain-of-thought prompting (i.e., GPT-4o with the ZS-CoT prompt). This underscores that facilitating a model's reasoning process can enhance overall performance. However, given our instruction to output a single word, a potential direction for future research to bridge the gap between model performance and user expectations could be to instead execute *silent* CoT reasoning to achieve better performance while still generating brief responses.

4.2. Locating All Plausible Regions of Focus

Models. We evaluate three models. Included is the state-of-the-art language grounding model, GLaMM [43]. We also benchmark two engineered solutions that facilitate the *reasoning process* by breaking the task into two simpler, sequential steps. The first relies on GPT-4o to generate description of the focus regions and GLaMM to locate the regions given the descriptions. The second prompts Molmo [12] to generate points (i.e., `x,y` coordinates) locating all of a question's focus regions and then feeds those as point prompts to SAM [28] to decode into segmentations.

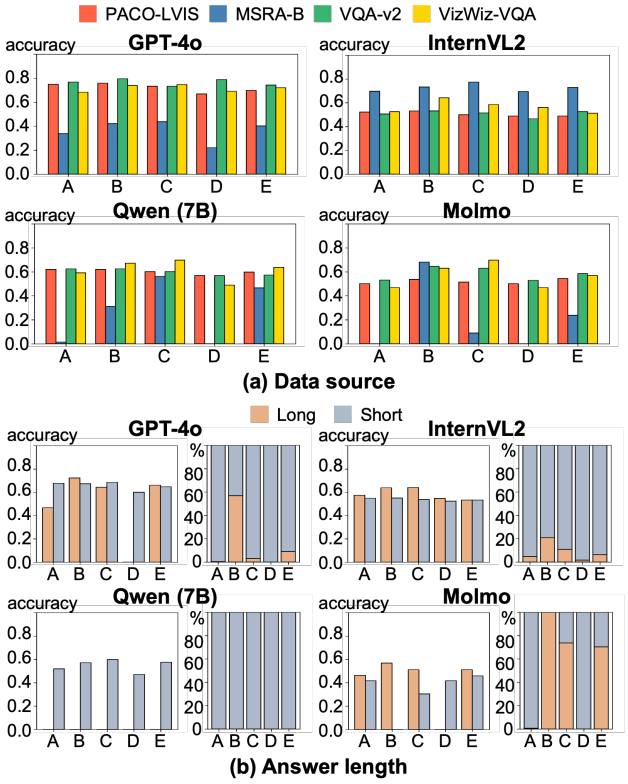

Figure 6. Fine-grained analysis of four models in recognizing focus ambiguity. Accuracy is reported based on data source and answer length. Percentages are provided relative to answer length. The five prompts are represented as follows: ZS (A), ZS-CoT (B), ZS-ECoT (C), FS (D), and FS-ECoT (E).

Prompts. For all models, we adopt the five prompting methods defined in Section 4.1, with minor modifications. For GLaMM, the problem definition instead indicates the task is to segment all focus regions. For ChatGPT-4o+GLaMM, we acquire descriptions of the focus regions from GPT-4o using the five prompts, and then acquire segmentations by passing each into a prompt for GLaMM specifying "Can you segment {description}?". For Molmo+SAM, we replace the aforementioned prompt's "segment" with "point", since Molmo generates points.

Evaluation Metrics. We employ three metrics. First is the standard metric for instance segmentation: *mAP*. We also employ *union IoU* and *max IoU* to analyze performance when models predict only one focus region. We calculate *union IoU* as the IoU between the predicted mask and the union of all focus regions to measure if the generated mask provides a semantic segmentation (rather than instance segmentations) capturing all focus regions. We calculate *max IoU* as the largest mIoU score between the predicted mask and each focus region, to see if the generated mask instead accurately captures a single focus region.

Overall Performance. Results are shown in **Table 3**. All models perform poorly, with a considerable performance discrepancy across them.

We attribute Molmo+SAM's leading performance (in terms of mAP) to its tendency to point to *multiple* regions, as reinforced by Molmo's nearly 100% positive rate observed for zero-shot settings in **Table 2** indicating its bias to predict question ambiguity. **Figure 7(b)** shows where it successfully grounds two small clocks. However, Molmo+SAM features much lower union IoU, which is because SAM can fail to segment the whole object with only point input, such as for a "drinking duck" where only the leg

Figure 7. Molmo+SAM zero-shot results. Questions and ground truth masks are shown in Figures 1 and 5, stars denote where Molmo points, and blue masks denote SAM's segmentations.

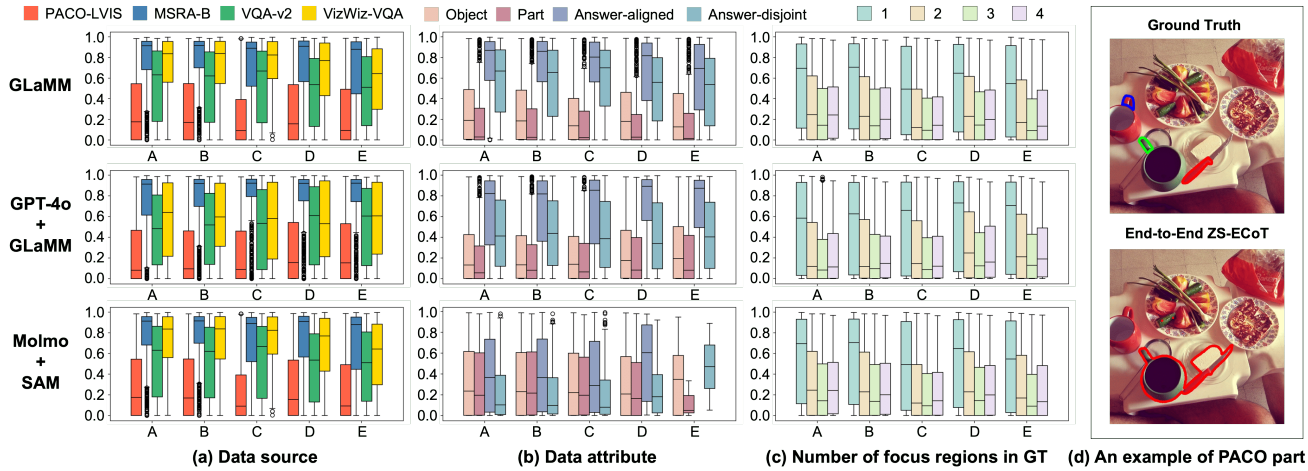

Figure 8. Fine-grained analysis on the performance of our three benchmarked approaches for question grounding using the five prompts, ZS (A), ZS-CoT (B), ZS-ECoT (C), FS (D), and FS-ECoT (E). (a) Union IoU scores of our four data sources. (b) Union IoU scores of object and part in PACO and the attribute based on question-answer question grounding alignment. (c) Union IoU scores of samples based on the number of focus regions. (d) An example of PACO parts grounded as focus regions in our dataset.

Approach	Prompt	mAP	union IoU	max IoU
GLaMM	ZS	13.01	**41.90**	**43.69**
	ZS-CoT	**13.24**	41.72	43.51
	ZS-ECoT	10.21	36.55	35.66
	FS	11.93	40.08	42.58
	FS-ECoT	10.29	37.01	39.21
GPT-4o+GLaMM	ZS	12.58	37.35	43.62
	ZS-CoT	13.04	37.99	44.78
	ZS-ECoT	13.76	38.24	43.39
	FS	**14.24**	**40.97**	**47.83**
	FS-ECoT	13.89	40.51	46.89
Molmo+SAM	ZS	23.9	**36.4**	44.6
	ZS-CoT	**24.3**	36.2	**45.4**
	ZS-ECoT	24.2	36.1	44.3
	FS	11.0	16.1	20.8
	FS-ECoT	-	-	-

Table 3. Performance of three models for focus ambiguity localization with respect to three metrics.

of the duck is segmented rather than the whole duck. Additionally, Molmo encounters errors when processing FS-ECoT due to the long context input, which results in it not able to make predictions for FS-ECoT (denoted by "–").

The other two models perform poorly for different reasons. While GLaMM supports multiple segmentation outputs, it consistently generates only a single mask. As for the GPT-4o-based model, its descriptions poorly correlated with the number of ground truth regions resulting in poor subsequent performance from GLaMM.

Fine-Grained Analysis. We finally perform fine-grained analysis, with all results shown in **Figure 8**.

First, we observe notable performance differences across dataset sources (**Figure 8a**). Almost all models with different prompt settings perform best on MSRA-B, followed by VizWiz-VQA and VQA-v2, and lastly PACO. We attribute better performance on MSRA-B to the ground truth often aligning with the most salient object, which simplifies localization. In contrast, PACO-based data often contains many focus regions occupying smaller areas, which can increase the segmentation difficulty.

Our fine-grained analysis also show models struggle to identify parts, when comparing models' performance in locating PACO's objects versus parts (**Figure 8(b)**), as exemplified in **Figure 8(d)**. This finding parallels progress in the broader computer vision community, where only relatively recently researchers have begun trying to segment parts.

Our results also underscore a correlation between performance and the number of focus regions, with worse performance when there are a greater number of focus regions (**Figure 8(c)**). This is exemplified in **Figure 7(a)**.

Finally, we found models perform consistently worse on examples where the question grounding differ from answer grounding. This is exemplified in **Figure 8(b)** and **Figures 7(c)** and **7(d)**.

5. Conclusion

We introduced the VQ-FocusAmbiguity dataset for evaluating models' abilities to acknowledge question ambiguity. Analysis reveals this dataset comes with unique challenges not explored in related VQA grounding datasets, particularly for answer grounding. Benchmarking reveals models struggle to recognize question ambiguity and locate all focus regions, underscoring the need for future research. We publicly-share the dataset to facilitate future progress.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[4] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280, 2019. 2

[5] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. 2

[6] Chongyan Chen, Samreen Anjum, and Danna Gurari. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15315–15325, 2023. 2, 3

[7] Chongyan Chen, Mengchen Liu, Noel Codella, Yunsheng Li, Lu Yuan, and Danna Gurari. Fully authentic visual question answering dataset from online communities. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024. 2, 3

[8] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020. 2

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5

[10] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024. 5

[11] Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. *arXiv preprint arXiv:1604.02125*, 2016. 2

[12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 5, 6, 7

[13] Deniz Engin and Yannis Avrithis. Zero-shot and few-shot video question answering with multi-modal prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2804–2810, 2023. 3

[14] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1811–1820, 2017. 2

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi 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*, pages 6904–6913, 2017. 3

[17] Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*, 2021. 2

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2, 3

[19] Danna Gurari and Kristen Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3511–3522, 2017. 2

[20] D. Gurari, K. He, B. Xiong, J. Zhang, M. Sameki, S. D. Jain, S. Sclaroff M. Betke, and K. Grauman. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation(s). *International Journal of Computer Vision*, 2017. 3

[21] Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). *International Journal of Computer Vision*, 126:714–730, 2018. 3

[22] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1, 2, 3

[23] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International jour-*

nal of multimedia information retrieval, 9(3):171–189, 2020. 2

[24] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 2

[25] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. Long-form answers to visual questions from blind and low vision people. *arXiv preprint arXiv:2408.06303*, 2024. 3

[26] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013. 2

[27] Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. *arXiv preprint arXiv:2310.14696*, 2023. 2

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 7

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[30] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[32] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 3

[33] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010. 2, 3

[34] Fariba Lotfi, Amin Beheshti, Helia Farhood, Matineh Pooshideh, Mansour Jamzad, and Hamid Beigy. Storytelling with image data: a systematic review and comparative analysis of methods and tools. *Algorithms*, 16(3):135, 2023. 2

[35] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3

[36] Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209, January 2021. 2

[37] Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14367–14388, 2023. 2

[38] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020. 2

[39] OpenAI. Gpt-4o system card, 2024. Accessed: 2024-09-01. 3, 5

[40] Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *arXiv preprint arXiv:2310.05861*, 2023. 2

[41] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 2

[42] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*, 2023. 2, 3

[43] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 7

[44] Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11953–11962, 2023. 2

[45] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[46] Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in vqa. *arXiv preprint arXiv:2211.07516*, 2022. 2

[47] Yu-Yun Tseng, Alexander Bell, and Danna Gurari. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *European Conference on Computer Vision*, pages 575–591. Springer, 2022. 3

[48] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[49] Mengting Wan and Julian McAuley. Modeling ambiguity,

subjectivity, and diverging viewpoints in opinion question answering systems. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 489–498. IEEE, 2016. 2

[50] Jianming Wang, Wei Deng, Yukuan Sun, Yuanyuan Li, Kai Wang, and Guanghao Jin. Twice opportunity knocks syntactic ambiguity: A visual question answering model with yes/no feedback. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 736–741. IEEE, 2019. 2

[51] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 2

[52] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[53] Chun-Ju Yang, Kristen Grauman, and Danna Gurari. Visual Question Answer Diversity. In *HCOMP*, pages 184–192, 2018. 2

[54] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3, 5

[55] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[56] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 2