

Acknowledging Focus Ambiguity in Visual Questions: Supplementary Material

This document supplements the main paper with more information about:

1. User Interface to Mitigate Ambiguity (Supplements Section 1)
2. Dataset collection (Supplements Section 3.1)
 - Image Sources for PACO-LVIS
 - Automated Candidate Question Generation
 - Annotation Rules
 - Annotation Task Design
 - Annotation Collection
 - Annotated Examples
3. Dataset analysis (Supplements Section 3.2)
 - Answer Groundings versus Question Groundings
4. Model Analysis (Supplements Section 4)
 - Model Details
 - Prompting Methods
 - Recognizing Questions with Focus Ambiguity
 - Locating All Plausible Regions of Focus

1. User Interface to Mitigate Ambiguity

This paper proposes a new directions for VQA by enabling the recognition of question ambiguity and localization of all focus regions. These can serve as valuable back-end capabilities to enable novel front-end human-computer interactions. For example, a front-end system could simply notify the user that ambiguity is detected, enabling the user to choose to try again with a modified question, image, or both. Alternatively, as exemplified in **Figure 1**, when ambiguity is detected, all plausible focus regions can be presented to the user for them to select one or more specific regions of interest through simple clicks.

2. Data Collection

2.1. Image Sources for PACO-LVIS

From preliminary analysis, we found that images with sufficient complexity for generating questions with focus ambiguity typically had either more than three segmented *object* instances or more than three segmented *part* instances. Therefore, our 2,272 samples sampled from the PACO-LVIS dataset were only those that met this complexity requirement.

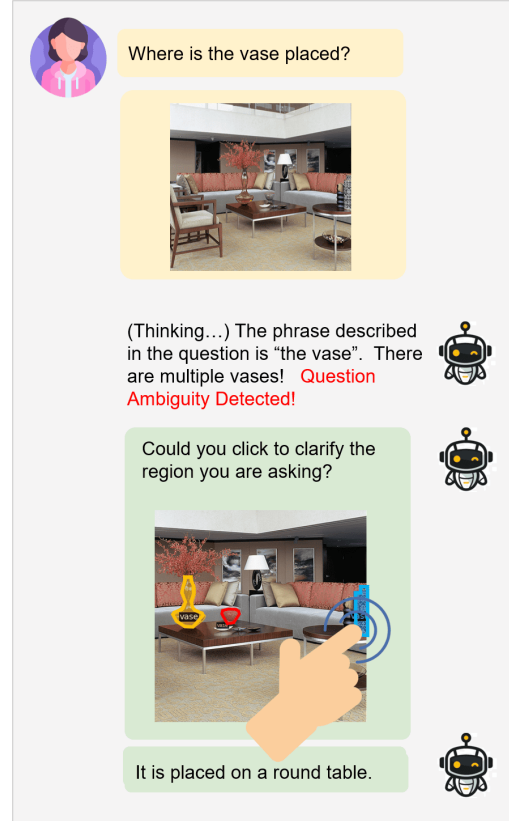


Figure 1. Example of a possible user interaction interface that could help mitigate ambiguity with our focus ambiguity recognition and localization tasks. Specifically, when ambiguity is detected, the user interface will first present all plausible focus regions in the image and then prompt users to select one or more specific regions through simple clicks.

2.2. Automated Candidate Question Generation

Question Candidate Generation for PACO-LVIS. We used GPT-4o to generate both a set of ambiguous questions that could focus on multiple regions in an image and a set of unambiguous questions that focus on a single region in an image. To achieve this, we fed four inputs to GPT-4o: instructions, the image, semantic labels within the 76 object or 200 part categories in PACO-LVIS for segmentations, and segmentation plots. To facilitate the model in crafting

questions we were seeking, we also provided both positive and negative in-context examples. These came from existing visual question answering datasets (i.e., VizWiz-VQA and VQAv2) as well as examples created by the authors that were provided as part of the instructions given to crowd-workers to help them author questions.

To facilitate generating a diversity of questions, we conveyed in the instructions that a diversity of questions is important and we employed the following three different types of prompts:

- **Default Prompt:** The input included the image, the definition of question ambiguity, step-by-step guidance which included things to avoid and to be careful about, format requirements (e.g., format the generated question in curly brackets), semantic labels for the segmentation, and both positive and negative examples.
- **Default + Segmentation Overlay:** In addition to the default prompt inputs, segmentation overlays were provided as supplementary image input. These overlays displayed all available segmentations on the original image using colored masks. Consequently, the input included both the original image and image with segmentations overlaid on the original image.
- **Default + Segmentation Overlay + Mirroring Requirements + Explanation:** This prompt built on the previous set-up by asking the model to also repeat the task requirements before generating questions and explain why the generated question satisfies the requirements.

We observed that the third prompt variation significantly improved question quality, likely due to the added clarity from reiterating instructions and additional reasoning process by providing explanations for the generated question. However, this approach was computationally intensive, requiring approximately 10 seconds per example.

Finally, we post-processed the responses from GPT-4o with regular expressions to extract the question from the curly brackets, mentioned in the format requirements. We publicly-share the code for generating automated candidate questions and post-processing at <https://focusambiguity.github.io/>.

Question Candidate Generation for MSRA-B. We create the question by randomly sampling from a question pool, which consists of the following variants of “What is this”: “What is this?”, “What is that?”, “What’s this?”, “What’s that?”, “What is this thing?”, “What is prominent?”, “What is in the foreground?”, “What is close to the camera?”, “What stands out?”, “What is at the front?”, and “What is featured up front?”.

2.3. Annotation Rules Design

We designed the annotation rules as follows: (1) one author identified five rules¹ by analyzing thousands of visual questions. Next, (2) three other authors independently annotated tens of diverse examples with these rules to identify all instances where their annotations differed, and then (3) the three authors refined the rules to prevent those differences going forward. The rules, provided alongside examples, clarified how they should be applied in practice by annotators during large-scale annotation collection.

2.4. Annotation Task Design

Annotation Task Design for PACO-LVIS. As mentioned in the main paper, we first asked the user to provide the question, and then select segmentations to which the question could be grounded. To collect the segmentations, we initially conducted a small-scale test with two independent annotators working under two different settings to examine the potential impact of bias from the user interface design: (1) all segmentations were initially unselected then clicking the mouse cursor on a region would select and clicking again will deselect it and (2) all segmentations were initially selected and then the user would deselect any irrelevant segmentations. The UI interface for setting (1) is shown in **Figure 2** and the UI interface for the setting (2) is shown in **Figure 3**. Both designs led to similar results (i.e., an exact match for 10 of 10 tested samples) while (1) costs an average of 0.85 minutes per example (1.7 minutes per HIT) in our pilot study and (2) was extremely time-intensive, taking over 30 minutes in the most demanding cases involving over 20 objects and 40 very small parts. Therefore, we proceeded with the large-scale collection with setting (1), with all segmentations initially unselected.

Annotation Task Design for Visual Questions. We created two UI interfaces for VQA-AnswerTherapy.

The first UI is for ambiguity label collection, and a screenshot is shown in **Figure 4**. It displays the image along with its corresponding question and all unique answers provided for that visual question. The users are then asked to identify the target entities described in the question by first identifying all entities and then selecting the correct entities, after which the annotator specifies whether the question is ambiguous.

The second UI is for locating the regions, and a screenshot is shown in **Figure 5**. We utilized the Supervisely software, as it supports segmentation tasks and grouping annotations. We provide the image and question to the users

¹The rules addressed cases include single or multiple entities, demonstrative pronouns, entire-image references, and questions with options provided.

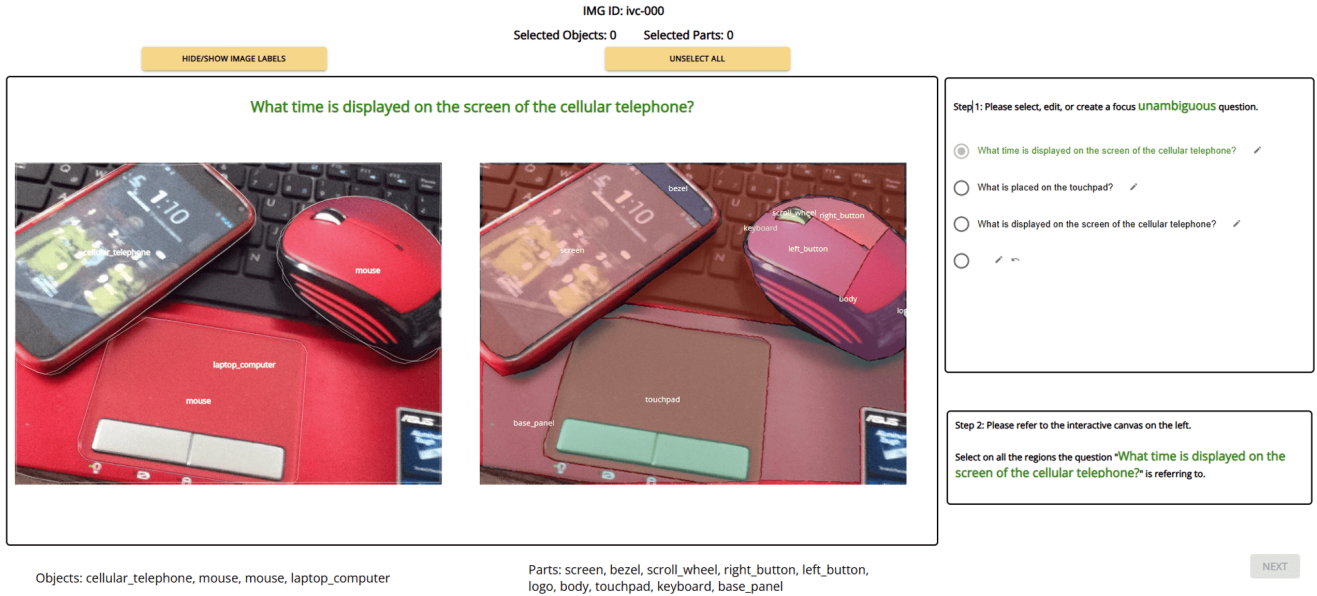


Figure 2. UI interface for collecting annotations for PACO-LVIS dataset with segmentations initialized unselected.

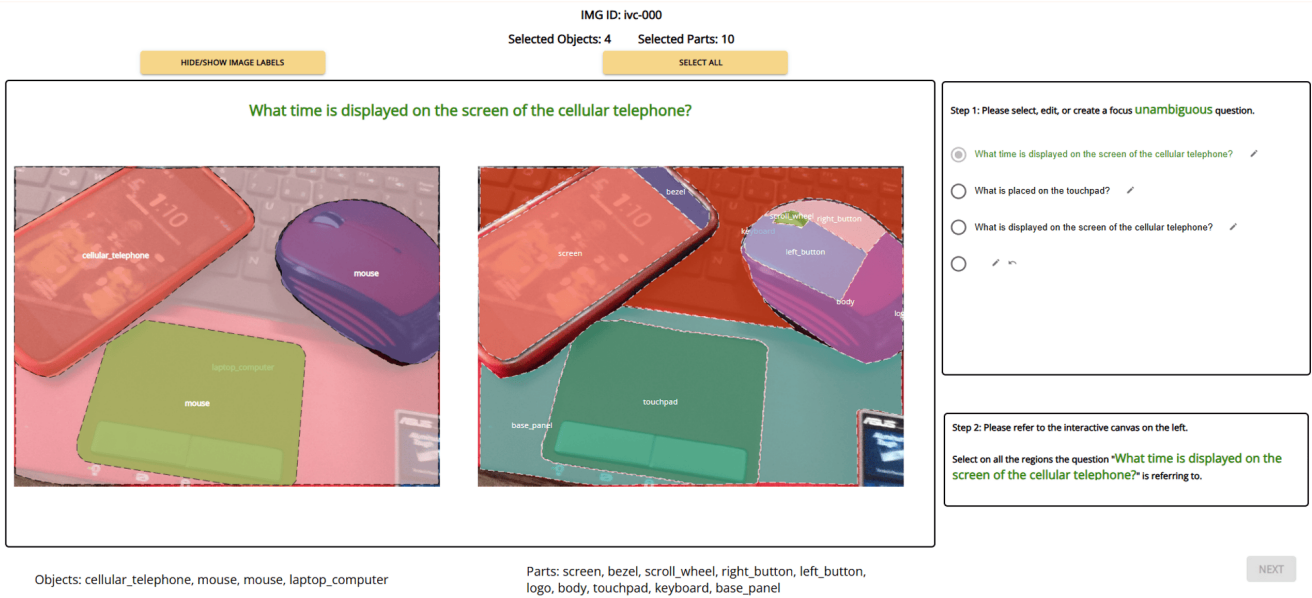


Figure 3. UI interface for collecting annotations for PACO-LVIS dataset with segmentations initialized selected.

without the answers to ensure the question groundings are independently done and not influenced by the answers.

2.5. Annotation Collection

Annotation Collection for PACO-LVIS. We took several steps to collect high-quality annotations. We hired five experienced crowd workers from Amazon Mechanical Turk to generate ambiguous and unambiguous questions and provide the question groundings for the questions who had previously collaborated with our team for at least three other

dataset creation efforts involving VQA and segmentation. These workers were based in the United States and had completed a minimum of 500 Human Intelligence Tasks (HITs) with an acceptance rate exceeding 95%. Each candidate worker received personalized training via a one-on-one Zoom session focused on our specific task requirements. We paid them \$0.5 per HIT, where each HIT requires creating two examples per image (1 ambiguous and 1 unambiguous), with a median of 0.85 minutes and a mean of 2.25 minutes per example. We also conducted both manual

DATA ANNOTATOR

LABELLED DATA

ANNOTATE DATA

Question

What is it?

Which object are we looking for? *

it

Referring expression? *

it

Noun chunks

Is the question ambiguous?

☒ Yes
 ☐ No

Notes

It here is pointing to multiple things. The answers arent necessari

NEXT

SUBMIT

Figure 4. UI interface for collecting the question ambiguity labels for the VQA-AnswerTherapy dataset.

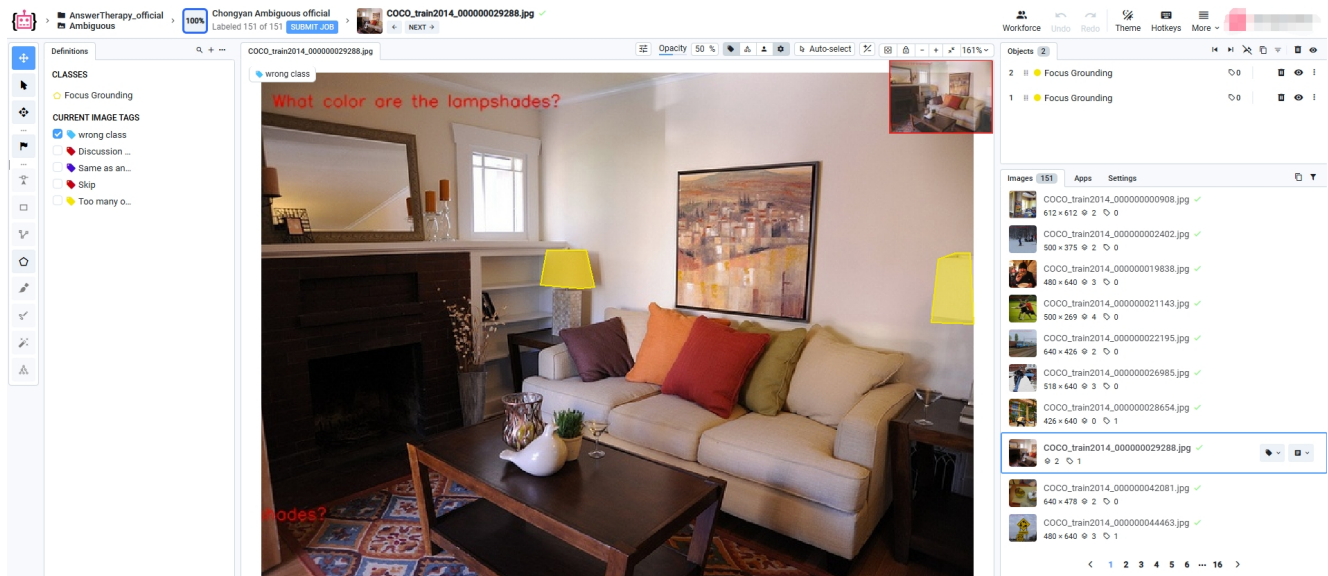


Figure 5. UI interface for collecting segmentation of regions described by the question for the VQA-AnswerTherapy dataset.

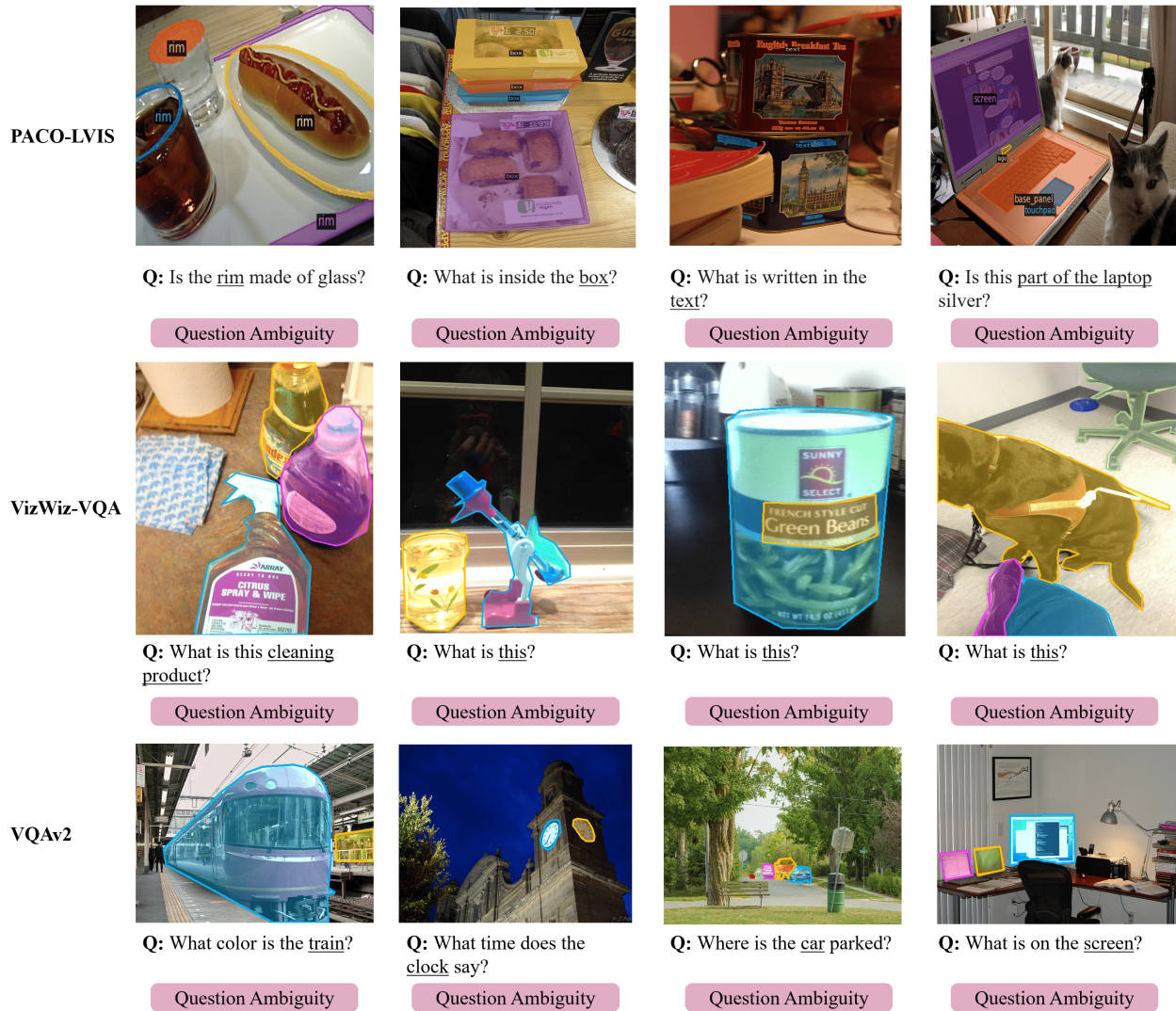


Figure 6. Examples of visual questions with focus ambiguity from three data sources.

and automated quality control mechanisms. For the manual quality control, we performed ongoing spotchecks through annotation collection and provided feedback to each of the workers as needed. For the automated quality control, we monitored time to task completion and the number of selected segmentations to identify potential outliers for manual inspection.

Annotation Collection for VQA-AnswerTherapy. One author annotated whether there was ambiguity in the reference of a phrase within the question. This process took one author about 3 weeks to finish with a minute or two to annotate a single image. For quality control, all edge cases are discussed between authors, and the 330 examples which have segmentations were verified by the other two authors.

Two authors collaboratively segmented the location of the question’s target phrase. Specifically, one author independently labeled ambiguous questions, while the other labeled unambiguous ones. They then switched roles to cross-check each other’s annotations. Discrepancies were discussed and resolved collaboratively. In total, it took approximately 15 hours to annotate 330 examples.

2.6. Annotated Examples

We show annotated examples of ambiguous questions along with the question groundings in **Figure 6** and examples of unambiguous questions along with the question groundings in **Figure 7**.






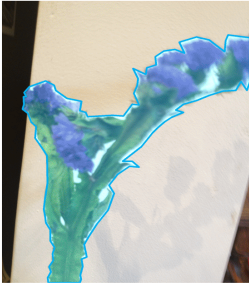










PACO-LVIS				
	Q: How many <u>buttons</u> are present on the TV set?	Q: What is written on the <u>lid of the top box</u> ?	Q: How many <u>remote controls</u> are displayed in the image?	Q: What is <u>the dog</u> sitting on?
	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity
VizWiz-VQA				
	Q: What is <u>that</u> ?	Q: What is <u>this</u> ?	Q: What am I looking at?	Q: What color is the <u>sink</u> ?
	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity
VQA v2				
	Q: What color is the <u>bench</u> ?	Q: What is behind the <u>woman on phone</u> ?	Q: Who is in the <u>middle of the road</u> ?	Q: What <u>brand of gas station</u> is shown in the picture?
	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity
MSRA-B				
	Q: What is prominent?	Q: What's stands out?	Q: What's <u>that</u> ?	Q: What's is close to the camera?
	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity	No Question Ambiguity

Figure 7. Examples of visual questions without focus ambiguity from four data sources.

3. Data Analysis

Answer Groundings versus Question Groundings. As discussed in the main paper, the answer grounding and question groundings can be different. To establish this, we manually reviewed all 330 examples' question and answer groundings to determine whether they match or dif-

fer. This is because all question groundings are annotated from scratch, and only considering IoU between question groundings and answer groundings might not provide an accurate evaluation. This is particularly true for small objects, where even slight boundary misalignments can significantly affect IoU scores. Additionally, question groundings and

answer groundings may refer to the same object but exhibit boundary misalignments due to differences in annotators’ interpretations. In occluded scenarios, annotators might define regions differently—for example, one might include the occluded portion of an object, while another might exclude it—resulting in mismatches despite semantically correct predictions.

Additional examples are shown in **Figure 8**. As shown, the answer grounding and question groundings can be different because the question is asking about the relationship between things, (e.g., “What is the person standing on?” when the focus is “person” while the answer is “floor”/“carpet”) or locations of the entities (e.g., “Where is the vase placed?”). It also happens when the focus is clearly about an item but the answer can be features of the item (e.g., “What does this say?” when the focus is the product and the answer grounding is about the text label on the product). It also happens when there are multiple possible foci, but the answer only mentions the one that is fully visible; e.g., the one in the center compared to those not in the center; the one taking a larger fraction of the image compared to those that are smaller; the one in the foreground compared to those in the background; the one that is readable compared to those that are blurry.

We provide the breakdown of the number of questions where question groundings and answer groundings match and are different with respect to the presence or absence of focus ambiguity (i.e., with multiple question groundings or single question grounding) and the presence or absence of answer ambiguity (i.e., with multiple answer groundings or single answer grounding) in **Table 1**. Qualitative examples are shown in **Figure 9**. This further supports the conclusions from **Figure 8** that the differences between question groundings and answer groundings may arise when the question pertains to relationships between the entities, as shown in **Figure 9 (d)** and **(e)** or locations of the entities, as shown in **Figure 9 (c)**.

4. Model Analysis

4.1. Model Details

We used for recognition task ChatGPT-4o, InternVL2-Llama3-76B, Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-72B-Instruct, and Molmo-7B-

	Single Answer G.		Multiple Answer G.	
	Match	Different	Match	Different
Single Focus G.	101	41	5	18
Multiple Focus G.	13	77	21	54

Table 1. Number of questions where question and answer groundings are matching and different, with respect to Single/Multiple Question Groundings and Single/Multiple Answer Groundings.



Figure 8. Examples of visual questions when the question grounding and answer groundings are different.

D-0924. The models we used for localization task are Molmo-7B-D-0924, GLaMM-FullScope, ChatGPT-4o, and SAM-ViT-h-4B.

4.2. Prompting Methods

Ablation Pilot Study for Prompts. To settle on the basic prompt (ZS), we first tested four variants of focus ambiguity definition on the GPT-4o model in zero-shot settings on all visual questions in VizWiz-VQA and VQAv2 for the recognition task as a pilot study. For these prompts, we substituted the text by modifying the definition of focus ambiguity in “You are a helpful assistant. You will be given an image and a question about the image. Your task is to predict whether the question is ambiguous or unambiguous based on the given image and the definition of focus ambiguity. Focus ambiguity in a visual question occurs when...(definition of focus ambiguity). Please only output “ambiguous” or “unambiguous”.”. The four different definitions are: (1) “if under-specified language can be grounded to multiple regions in the image” (F1: 78, Acc: 76), (2) “when the question does not clearly indicate which



Q: What is prominent?
F: Ferret A: Ferret

(a) No Question Ambiguity
No Answer Ambiguity
Matching



Q: What color are the rugs?
F: rugs A: beige
A: green

(b) No Question Ambiguity
Answer Ambiguity
Matching



Q: Where is the bird?
F: bird 1 A: on bird house
F: bird 2 A: bird house

(c) Question Ambiguity
No Answer Ambiguity
Different



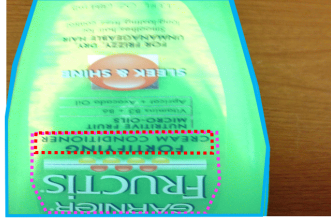
Q: What is the food on?
F: food (pizza) A: plate
F: food (salad) A: dishes
A: table

(d) Question Ambiguity
Answer Ambiguity
Different



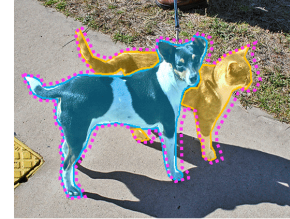
Q: What is behind the woman on phone?
F: Woman A: mountains
on phone A: mountain

(e) No Question Ambiguity
No Answer Ambiguity
Different



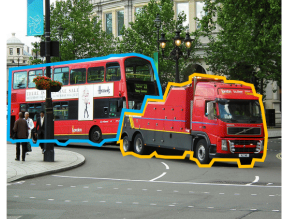
Q: What does this say?
F: this A: fortifying cream conditioner
A: garnier fructis fortifying cream conditioner

(f) No Question Ambiguity
Answer Ambiguity
Different



Q: What animal is in the scene?
F: animal A: cat and dog
F: animal A: dog and cat

(g) Question Ambiguity
No Answer Ambiguity
Matching



Q: What type of vehicle is here?
F: vehicle A: tow truck
F: vehicle A: bus

(h) Question Ambiguity
Answer Ambiguity
Matching

Figure 9. Examples of visual questions are shown with question groundings and answer groundings overlaid on images from VQAv2, VizWiz-VQA, PACO-LVIS, and MSRA-B. These examples illustrate various combinations, including the presence or absence of focus ambiguity, the presence or absence of answer ambiguity, and whether the focus and answer groundings match or differ.

part of the image it specifies” (F1: 76, Acc:74), (3) “if the question refers to multiple regions in the image.” (F1: 74, Acc:73) (4) and “if there are multiple regions in the image each satisfying the question’s constraints” (F1: 77, Acc: 75). The phrasing “if under-specified language can be grounded to multiple regions in the image” proved to be the most effective and thus we selected this definition.

We then tested whether introducing the term “focus ambiguity” to the model in the prompting as opposed to simply providing the conditions “An visual question is ambiguous if It is unambiguous if...”. The results show a boost to an F1 score of 80 for the AnswerTherapy source. Therefore,

we used this last format. The final prompt for zero-shot setting is shown in **Figure 10**.

Details of Prompts for Each Experiment. In model benchmarking, all models are tested with five prompting methods, zero-shot (ZS), zero-shot chain of thought (ZS-CoT), zero-shot enhanced chain of thought (ZS-ECoT), few-shot (FS), and few-shot enhanced chain of thought (FS-CoT). These are illustrated in **Figure 10, 11, 12, 13, and 14**.

For *focus ambiguity recognition*, we prompted GPT-4o, InternVL2, Qwen2, and Molmo using these five prompts,

“You are a helpful assistant. You will be given an image and a question about the image. Your task is to predict whether the question is ambiguous or unambiguous based on the given image.

A question is ambiguous if under-specified language in questions can be grounded to multiple regions in the images, leading to uncertainty about the intended focus. Specifically, ambiguity arises when multiple regions in the image could each satisfy all the constraints of the question, but the question fails to clarify which one it refers to (e.g., by not using the plural form or failing to specify a region).

A question is unambiguous if it clearly identifies or specifies the part of the image it refers to, leaving no uncertainty about the intended focus. If the image contains only one object or region that satisfy all the constraints of the question, and there is no possibility of multiple interpretations due to its simplicity, the question is considered unambiguous, even if the question lacks specificity or the phrasing is vague (e.g., ‘What is the color of the object?’ when there is only one object shown in the image); If the question explicitly refers to multiple regions using the plural form or clear descriptors, the question is also considered unambiguous.

Please only output “ambiguous” or “unambiguous”.

Figure 10. The prompt for Zero-Shot setting.

“(omitted, same as zero-shot setting...) ”

Please think step by step first, and then predict whether it is ambiguous or not.

Please only output “ambiguous” or “unambiguous”.

Figure 11. The prompt for ZS-CoT setting.

“(omitted, same as zero-shot setting...) ”

Please follow these steps to think step by step first, and then predict whether it is ambiguous or not:

1. Understand the Image: Carefully analyze and fully comprehend the given image.

2. Understand the question: Carefully analyze and fully comprehend the question.

3. Find the regions: Find all regions in the image that could each satisfy all the constraints and contain all the necessary information to answer the question.

4. Predict: Predict whether the question is ambiguous or not. If there are multiple regions found and the question does not use plural form, it is an ambiguous question. Otherwise, it is an unambiguous question.

Please only output “ambiguous” or “unambiguous”.

Figure 12. The prompt for ZS-ECOT setting.

GPT-4o only “Use a word or a phrase to describe each region. Please only output the descriptions of these regions and use commas to divide descriptions of regions if there are multiple regions.” We then prompt GLaMM using the each generated description embedded in “Can you segment {description}?” to acquire segmentation masks.

For *Molmo+SAM localization*, we used the same five prompts with the only difference being replacing the “segment” as “point” for Molmo generate points. The generated points’ coordinates are fed into SAM to generate the segmentations.

Discussion about Prompt Strategies. As shown in the main paper in section 4.1, when comparing different prompt types, we found that *across all models, ZS-CoT and ZS-ECOT consistently lead to better performance*. This is because, CoT is better than without CoT (not surprising); and surprisingly, ZS is better than FS across all models, except for Molmo (7B), in our setting.

We attribute this unexpected outcome to three interconnected factors: The few shot setting has (1) too limited (only two shots) and (2) challenging examples, (3) without additional explanations along with the examples. We intentionally selected two challenging examples for FS, including one unambiguous question with answer ambiguity, as shown in **Figure 13**. The original intention was to encourage models to differentiate clearly between question and answer ambiguity, similar to how contrastive learning and humans benefit from hard samples. However, with only two shots, it turned out to only confuse models and led to over-

respectively, to acquire classification results.

For *end-to-end region localization*, we prompted GLaMM by adjusting these five prompts to “segment each of the regions in one mask.” For example, for step 4 of the example 1 of FS-ECOT, we adjusted it to “Predict segmentation: Predict three segmentation masks separately - segmentation of the blue kite, segmentation of the red kite, and segmentation of the yellow kite.”

For *ChatGPT-4o+GLaMM localization*, we prompted GPT-4o using these five methods to acquire descriptions of the regions. We add additional formatting constraints let

“(omitted, same as zero-shot setting...) ”

Here are two examples with the images described for you:

Example 1:

Question: What color is the kite?

Image description: The image shows a group of people in a park with picnic tables. One table has some food on it and there are two women sitting at the table. Other people are standing. One young boy is holding a blue kite in his hand, another child is picking up a blue kite from the ground, and one man is running and flying a yellow kite.

Prediction: ambiguous

Example 2:

Question: Where is the man holding the apple?

Image description: The image shows a man wearing a blue shirt sitting at a dining table and a little girl sitting on the floor beside him. The setting is likely a dining room. There is a window behind the dining table with brown curtains. The man is holding an apple in his hand. The little girl is holding a toy in her hand. A brown dog is laying on the floor looking at the girl.

Prediction: unambiguous

Below is the task for you to make prediction based on the given image. Please only output “ambiguous” or “unambiguous”.

Figure 13. The prompt for FS setting.

Model size	ZS	ZS-CoT	ZS-ECOT	FS	FS-ECOT
72B	58.7	60.5	64.7	57.7	64.7
7B	57.2	63.8	65.5	53.6	59.0
3B	44.6	44.6	45.0	44.3	44.3

Table 2. Performance of Qwen2.5-VL model series in three model sizes, 72B, 7B, and 3B, presented in accuracy.

fitting towards ambiguous cases, possibly because the unambiguous case is too hard for model to understand without first learning from simpler, more basic cases, without additional explanation. We attribute this to the model exhibiting overfitting tendencies, with only the ambiguous case understood, evidenced by a significantly increased positive rate in FS for all models.

In contrast, FS-ECOT includes step-by-step explanations, significantly boosting performance with the same two hard examples. Given that most models are sensitive to prompt types, as shown in Table 2 in the main paper, we encourage future research to further explore prompt strategies, starting with basic and simple examples, and explore in-context fine-tuning/expanding context windows/integrating multi-image inputs to enable a richer set of few-shot examples, and provide explicit reasoning steps (e.g., ECOT) to help models to understand examples.

4.3. Recognizing Questions with Focus Ambiguity

Different Sizes of Qwen2.5-VL. We benchmarked Qwen2.5-VL across model sizes ranging from 72B to 3B,

as shown in Table 2. The 72B model achieves the best performance in ZS, FS, and FS-ECOT settings, while the 7B model performs best in ZS-CoT and ZS-ECOT. Notably, the 7B model with ZS-ECOT yields the highest overall performance across all sizes and prompting strategies, so we report its results in the main paper. In contrast, the 3B model consistently underperforms, which we attribute to its limited reasoning capabilities due to its significantly smaller LLM size.

Effect of Fine-tuning on Qwen2.5-VL. To understand whether fine-tuning on our dataset can improve recognizing question ambiguity, we fine-tuned Qwen2.5-7B improves accuracy by 1.6 percentage points to 58.8% (ZS), matching the larger Qwen2.5-72B (ZS). While we expect similar boosts for other models, overall performance remains low.

Large reasoning model. we conducted an ablation study with ChatGPT-o3 ($\geq 200B$). The results, (Table 3), reinforce that existing models struggle to recognize question ambiguity.

Ablation Study on Input Types for ChatGPT-4o. We conducted two ablation studies to assess the impact of different input types on ChatGPT-4o’s ability to recognize questions with focus ambiguity.

Question-only input. Given that existing models often exhibit strong language priors and may overlook visual input, a common issue in the vision-and-language domain,

“(omitted, same as zero-shot setting...)”

Please follow these steps to think step by step first, and then predict whether it is ambiguous or not:

1. Understand the Image: Carefully analyze and fully comprehend the given image.
2. Understand the question: Carefully analyze and fully comprehend the question.
3. Find the regions: Find all regions in the image that could each satisfy all the constraints and contain all the necessary information to answer the question.
4. Predict: Predict whether the question is ambiguous or not. If there are multiple regions found and the question does not use plural form, it is an ambiguous question. Otherwise, it is an unambiguous question. Please only output “ambiguous” or “unambiguous”.

Here are two examples with the images described for you:

Example 1 [INSERT SAME EXAMPLE 1 AS FS SETTING]

- Step 1. Understand the Image: Understand the given image as described.
- Step 2. Understand the question: The question is about the color of the object, the kite, in the image.
- Step 3. Find the regions: The three regions of the three kites are the regions that can each satisfy all the constraints and contain all the necessary information to answer the question.
- Step 4. Predict: There are three regions found — multiple regions make this question ambiguous. Therefore, the prediction is “ambiguous”.

Example 2: [INSERT SAME EXAMPLE 2 AS FS SETTING]

- Step 1. Understand the Image: Understand the given image as described.
- Step 2. Understand the question: The question is about where the man holding the apple is.
- Step 3. Find the regions: In the image we can only find one man holding the apple, which satisfies all the constraints and contains all the necessary information to answer the question.
- Step 4. Predict: There is only one region found. Therefore, the question is “unambiguous”.

Below is the task for you to make prediction based on the given image. Please follow the steps to predict whether the question is ambiguous or not. Please only output “ambiguous” or “unambiguous”.

Figure 14. The prompt for FS-ECOT setting.

we evaluated ChatGPT-4o using only the question text. As shown in Table 3, performance drops significantly across all prompts, with decreases ranging from 15.2% to 25.9%, highlighting the critical role of image input.

Caption as image substitute. In our few-shot setting, example inputs consist of a question paired with a textual description of an image, while the actual task involves answering a question given an image, due to the model’s current limitation in processing multiple images. To explore this modality mismatch, we benchmarked few-shot ChatGPT-4o using its own generated captions in place of image inputs at inference time. The results are comparable (Table 3); for example, few-shot accuracy increased by 1.3%, while FS-ECOT decreased by 2.1%.

	ZS	ZS-CoT	ZS-ECOT	FS	FS-ECOT
ChatGPT-o3	63.0	66.6	68.1	64.7	70.9
ChatGPT-4o (Q only)	44.2	45.9	42.5	44.8	44.9
ChatGPT-4o (ImgCap)	-	-	-	61.3	62.8

Table 3. Recognition (accuracy) results.

Qualitative Results Additional qualitative examples of results from the GPT-4o and InternVL2 models are shown in Figure 15. These exemplify our findings from our analysis of 40 random examples, with 20 from AnswerTherapy and 20 from PACO. Most of the time both GPT-4o and InternVL2 focus on one valid region from multiple options. Most answers don’t include clear specifications of the question region (example 1), while a few do (example 2). In rare cases, the answer encompasses all potential regions with clear specifications of each question region (example 3). None of the tested samples were followed up by a clarification question such as “Can you clarify which car you are asking about?”

Success Case of GPT-4o. We investigate the best performing model, GPT-4o, for the recognition task. We attribute GPT-4o’s advantage to its bigger size compared to other models as well as its extensive training with human alignment and real-world feedback, where ambiguity is likely common. To further explore GPT-4o’s success, we present qualitative results in Figure 17, showing samples

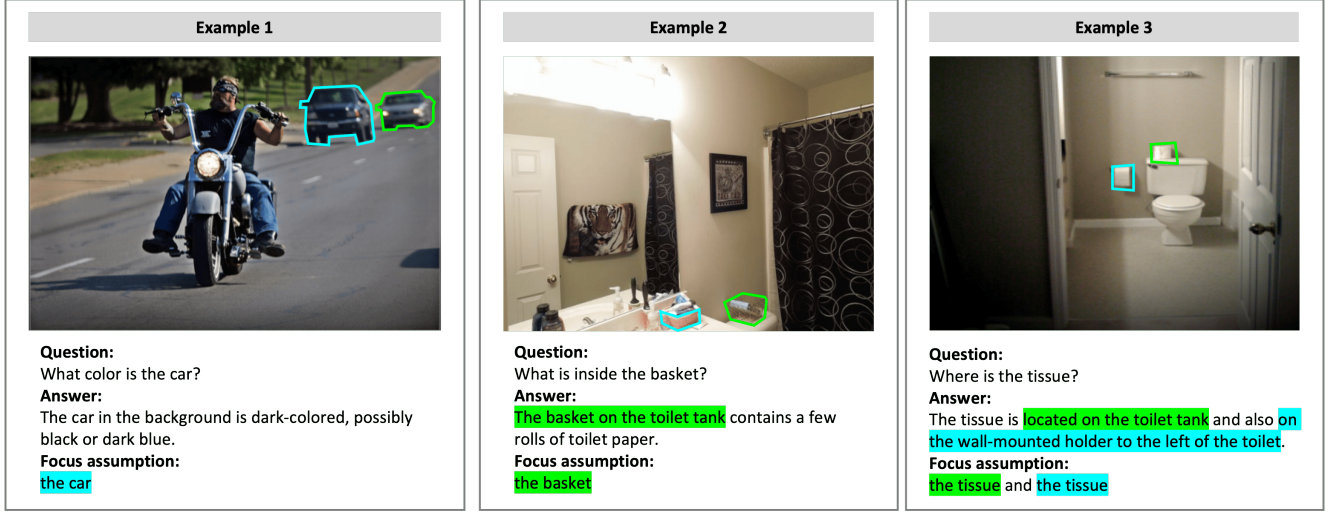


Figure 15. Examples of visual question answering for ambiguous questions in preliminary experiments. The coded colors highlight the specifications of the question regions in the answer and the question assumptions that we observe via the answers. Example 1 exemplifies the answer that targets one question region yet without specifying which region it is in the answer. Example 2 exemplifies the answer that targets one question region and specifies the region in the answer. Example 3 exemplifies the answer that encompasses all question regions and specifies all regions in the answer.

Example 1	Ground Truth	GLaMM [ZS] Prediction	GPT-4o + GLaMM [ZS] Prediction
Question: What is in the bowl?	Focus ambiguity: Ambiguous Focus region(s): the bowl the bowl the bowl the bowl	Focus ambiguity: Unambiguous	Focus ambiguity: Unambiguous Step 1 focus region description(s): eggs
Example 2	Ground Truth	GLaMM [FS] Prediction	GPT-4o + GLaMM [FS] Prediction
Question: What is on the plate?	Focus ambiguity: Unambiguous Focus region(s): the plate	Focus ambiguity: Unambiguous	Focus ambiguity: Ambiguous Step 1 focus region description(s): grilled sausage with onions and bread fries green relish in a small cup
Example 3	Ground Truth	GLaMM [ZS-CoT] Prediction	GPT-4o + GLaMM [ZS-CoT] Prediction
Question: What kind of crust is on the pizza?	Focus ambiguity: Unambiguous Focus region(s): the crust	Focus ambiguity: Unambiguous	Focus ambiguity: Unambiguous Step 1 focus region description(s): the edge of the pizza showing the crust

Figure 16. Qualitative results in GLaMM and ChatGPT-4o+GLaMM, illustrating various challenges in the localization task.

where GPT-4o succeeds while other models fail to recognize the focus ambiguity. We found that GPT-4o is superior in identifying unambiguous samples compared to other models, especially when other models also overlook constraints in the question; e.g., for “What color is the largest microwave?” and when the question uses a plural form, e.g., “What items are on the plates?”

Failure Case of GPT-4o. We then qualitatively review GPT-4o’s prediction errors and found it often struggles

when there is question ambiguity while the image prominently features a salient entity, as exemplified in **Figure 18(a)**; the model tends to interpret the scenario as unambiguous, ignoring other plausible entities that are less prominent. The model can also mistakenly deem unambiguous questions as ambiguous by overlooking constraints in the question; e.g., for “What is the title of the book with a blue cover in the center of the image?” only one of multiple books satisfies the constraint of having a blue cover. Other cases include (1) the question uses a plural form (2)

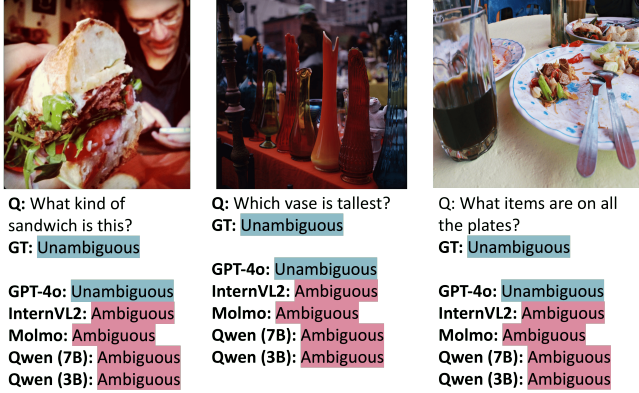


Figure 17. Qualitative results for examples where GPT-4o succeeds while other models fail.

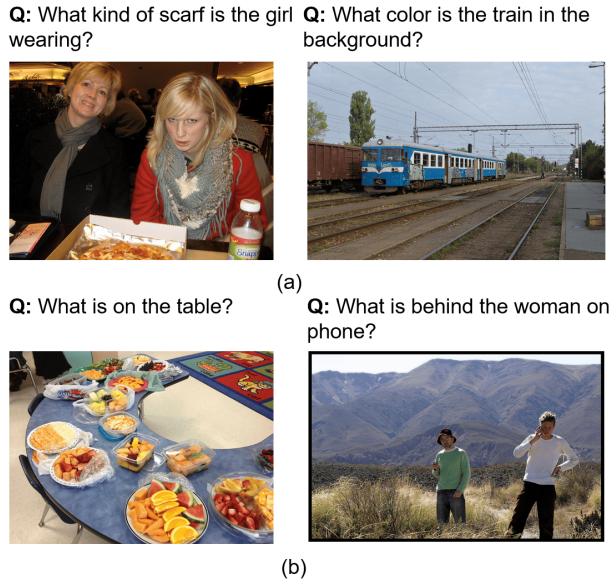


Figure 18. Qualitative results for GPT-4o, where (a) presents ambiguous question but GPT-4o predicted it as unambiguous and (b) presents unambiguous question but GPT-4o predicted it as ambiguous.

the question is unambiguous but the answer involves ambiguity, the model often provides ambiguous predictions. For example, in the question “What is on the table?”, if there is only one table but multiple types of food on it, as shown in **Figure 18(b)** in the left image, GPT-4o fails by predicting it as ambiguous.

Undecided Rate. The undecided rate indicates how often a model abstains to make a focus ambiguity prediction. We expected this rate to be inversely correlated with model performance, but surprisingly, this was not the case. For example, CoT, with better performance, often results in higher undecided rates. Another example is the worst-performing

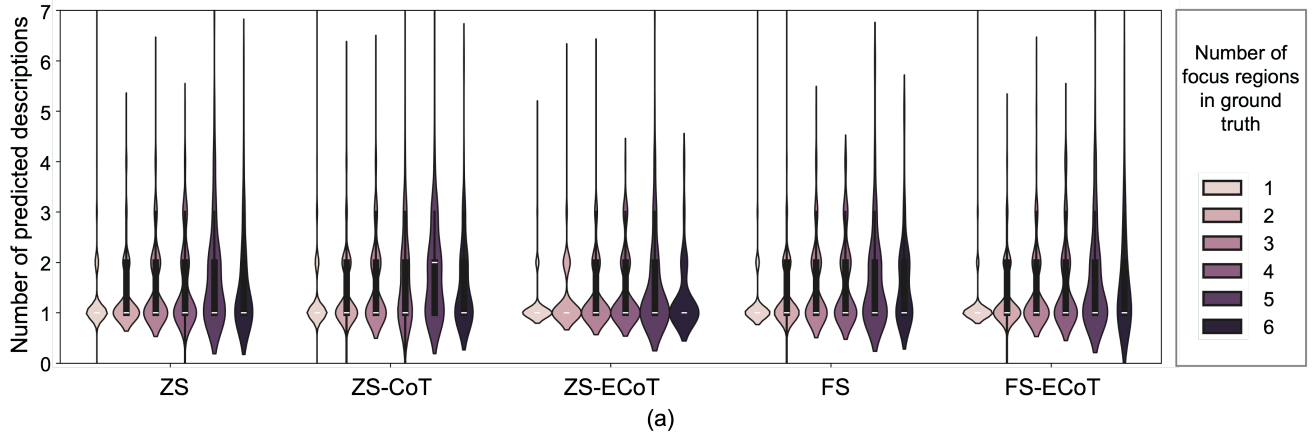
model, Qwen (3B), has the lowest undecided rate.

We suspect that models with stronger reasoning skills might realize and acknowledge its limitation rather than being over-confident and hallucinating; i.e. known unknowns vs. unknown unknowns. From qualitative results, we observe that undecided responses typically include phrases like “I’m unable to fulfill this request” and “I need more information”.

4.4. Locating All Plausible Regions of Focus

Challenges in Locating Question Regions. The *GLaMM approach*, *ChatGPT+GLaMM*, and *Molmo+SAM* all perform poorly in question focus localization. From the quantitative results in Section 4.2, we discover that the models struggle to localize regions in ambiguous questions (i.e., multiple groundings); regions in PACO-LVIS, especially those that are parts instead of objects in the dataset; and regions that don’t match answer groundings. To highlight these challenges and supplement the qualitative results for Molmo+SAM presented in the main paper, we provide additional qualitative results in **Figure 16** for the other two localization methods. Example 1 illustrates a case where both the *GLaMM approach* and the *ChatGPT4o+GLaMM approach* locate only the largest plausible focus region and miss the other regions. Example 2 illustrates a case where the *GLaMM approach* grounds both the question region and answer regions, while *ChatGPT4o+GLaMM* generates only the answers in step 1 and thus only grounds answers. Example 3 illustrates a case where the *GLaMM approach* localizes the question region relatively well, and step 1 in the *ChatGPT4o+GLaMM* generates a correct description but completely misses the described area in step 2. From these examples, we discover that both approaches demonstrate confusion between question groundings and answer groundings. Also, in *ChatGPT+GLaMM*, challenges can occur in both steps.

Analysis on Generated Descriptions in ChatGPT+GLaMM Approach. For the question focus grounding task, we introduce the *ChatGPT+GLaMM approach* to compensate for the poor performance of GLaMM in generating multiple regions in one answer. However, we did not observe an obvious improvement in the overall performance of the ChatGPT+GLaMM approach. We further break down the results of the first step, *describe*, and the second step, *localize*, to highlight the bottleneck of the task. In **Figure 19(a)**, we present the distribution of the predicted number of descriptions in the first step by the number of question focus regions in ground truth. We discovered that the overall segmentation performance is poor, and also the number of descriptions generated in the first step doesn’t have a strong correlation to the number of ground truth regions. In **Figure 19(b)**,



Example visual question: Is there a logo on the remote?

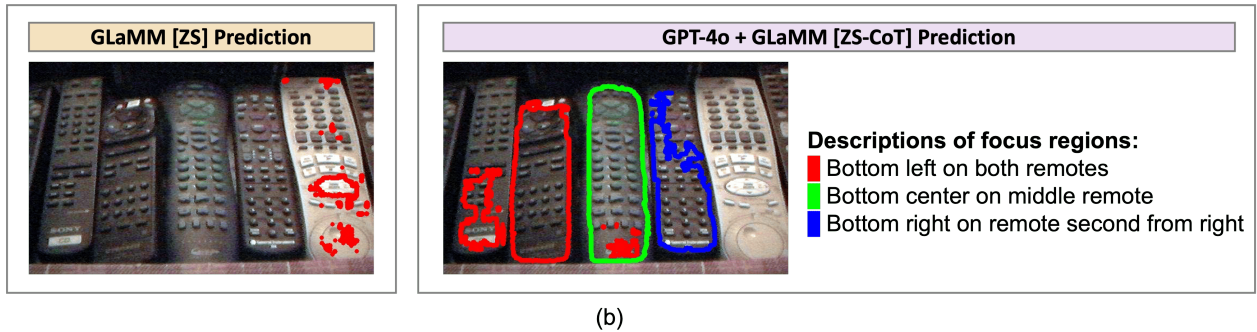


Figure 19. Analysis on the results of the prediction from GPT-4o in the GPT-4o + GLaMM approach for locating question focus regions. (a) Distribution of the predicted number of descriptions in the first step by the number of question focus regions in ground truth. (b) Example of multiple descriptions generated in step 1 and their grounding results compared to results of the *end-to-end* approach.

we observe that the number of described question focus regions increased in the ChatGPT+GLaMM approach compared to the end-to-end GLaMM approach. However, we can still see that the model suffers from not generating a clear description for every region. We suspect that the poor performance in the first step, *describe*, might impact the performance of the ChatGPT+GLaMM approach and thus reduce the improvement in its overall performance.

References