

VQA Therapy: Exploring Answer Differences by Visually Grounding Answers: Supplementary Material

1. Supplementary Material

This document supplements the main paper with more information about:

- Dataset collection (Supplements Section 3.1)
 - Method for hiring expert crowdworkers (Supplements Section 3.1)
 - Annotation task interface (Supplements Section 3.1)
 - Method for reviewing work from crowdworkers (Supplements Section 3.1)
- Dataset analysis (Supplements Section 3.2)
 - Incorrect answer (Supplements Section 3.2)
 - No polygon and multiple polygons (Supplements Section 3.2)
 - Grounding agreement (Supplements Section 3.2)
 - Reconciling redundant annotations (Supplements Section 3.2)
 - Four grounding relationships (Supplements Section 3.2)
 - Most common answers (Supplements Section 3.2)
- Algorithm benchmarking (Supplements Section 4)
 - Experimental details for Single Answer Grounding Challenge (Supplements Section 4.1)
 - Experimental details for Answer(s) Grounding Challenge (Supplements Section 4.2)
 - Performance of three models for Answer(s) Grounding Challenge with IoU-PQ metric (Supplements Section 4.2)
 - Answer(s) Grounding Challenge: qualitative results for model benchmarking (Supplements Section 4.2)

I. Dataset Collection

I.1. Method for Hiring Expert Crowd Workers.

We hired 20 workers who completed our one-on-one zoom training, passed our multiple qualification criteria, and consistently generated high-quality results. We limited the number of workers on our task to prioritize collecting *high-quality* annotations over the *efficiency* that would come with having more workers; i.e., it is easier to track the performance of fewer workers. We gave our 20 workers our contact information so that they could send any questions about the tasks and receive feedback quickly.

We paid above the US federal minimum wage to simultaneously support ethical data collection and encourage workers to create higher-quality results. Our average hourly wage was 9.64 dollars/hour. This rate is derived using the median time it took to annotate the 1,000 HITs collected in our pilot study (i.e., 2.49 minutes per HIT) with the amount we paid per HIT (i.e., 0.4 dollars/HIT).

I.2. Annotation Task Interface.

We show a screenshot of the crowdsourcing instructions in Figure 1 and the interface to collect annotations in Figure 2. The link to this code is available at <https://github.com/CCYChongyanChen/VQATheryCrowdsourcing/>.

I.3. Method for Reviewing Work from Crowdworkers.

In the first three days crowdworkers worked for us¹, we conducted highly interactive quality control. We conducted at least three inspections for each worker and gave them feedback continually. Each time, we viewed ten random HITs from each worker, provided each worker feedback if needed, and answered any questions by email or zoom. After the first time of review, 12 out of 20 workers passed our inspection without any issues. After the second time of review, 18 out of 20 workers passed our inspection without any issues. After the third time of review, all 20 workers demonstrated mastery of our task. We continued to monitor work from the eight workers who didn't work perfectly in the first time to ensure high-quality results.

¹The data collection process lasted for 26 days.

Hide / Show Instructions

Main Task

MOTIVATION

Our goal is to help blind people learn about their surroundings.
We aim to build an intelligent system that can automatically locate regions in images that are of interest to blind photographers.

TASK

In this task, you will see images paired with questions that were submitted by people who are blind and the answers are provided by multiple people. For each image-question pair, we collected answers from multiple people and so sometimes ended up with multiple different answers.

We will present to you three image-question pairs. For each image-question pair, we will ask you to review multiple answers and complete the following three steps for each answer.

- (1) Step 1: Is the answer correct?
- (2) Step 2: How many polygons are needed to locate the region that the answer is referring to?
- (3) Step 3: Draw one polygon to locate the region that the answer is referring to by clicking on the image.

Once you have completed the **answer** for an image-question pair, you will be allowed to proceed to the next answer. To go to the next answer, click the angle brackets ">" at the left bottom of the page.

Once you have completed the three steps for **all answers** for an image-question pair, you will be allowed to proceed to next image-question pair from the three image-question pairs.

To go to the next image-question pair, click the button "next image" at the right bottom of the page.

Once you have completed for **all image-question pairs**, you will be allowed to submit the HIT.

Step 1: Is the answer correct?

[▶ See details and examples](#)

If your answer is "Yes" to step 1, please go to step 2. Otherwise, click the angle brackets ">" at the left bottom of the page.

Step 2: How many polygons are needed to locate the region that the answer is referring to?

Zero : No polygon is needed. This is the situation when the **answer cannot be located** in the image.

[▶ See details and examples](#)

One: Just one polygon is needed. This is the situation when the answer is referring to a **single region** or multiple **connected** regions.

[▶ See details and examples](#)

More than one: Multiple polygons are needed. This is the situation when the answer is referring to **multiple disconnected regions**.

[▶ See details and examples](#)

If your answer is "one polygon" to step 2, please go to step 3. Otherwise, click the angle brackets ">" at the left bottom of the page.

Step 3: Draw one polygon to locate the region that the answer is referring to by clicking on the image.

Option (a): If the answer **has been** located in one of the previously drawn regions, select that region.

[▶ See details and examples](#)

Option (b): If the answer **has not been** located in one of the previously drawn regions, draw ONE polygon to locate the region that the answer is referring to following these instructions:

- **To draw**: Click the image to draw points one by one around the targeted region to form a polygon. No drag operation is needed.
- **To finish**: Click the first point again (the polygon will turn purple when your cursor is on the first point you draw). Or press keyboard shortcut 'Enter'.
- **To undo**: Click the Undo button. Or press keyboard shortcut 'Ctrl+Z'.
- **To clear**: Click the 'Clear' button.

[▶ See details and examples](#)

[▶ See details and examples](#)

NOTE

- Reminder: You will complete steps 1-3 for each answer to 3 question-based images in this HIT.
- Please do not refresh the webpage once you have started working, as you will lose all your work and have to start from the beginning.
- If you have any questions, please contact us at [\[redacted\]](#) If you wish us to notify you when we release new HITs, you can leave your email in the comment box when you submit the HITs. The comment box is optional, feel free to leave it blank.

Hide


You can see this information anytime by clicking "Hide / Show Details" button above.

Figure 1: Instructions for our annotation task.

Image 1 Image 2 Image 3

Question: What is this?

Unprocessed Answer(s): spoon, plate spoon



Clear Undo Select the whole image

Please read the question and answer about the image shown on the left. Then complete the 3 steps below.
We review the results. If you do not follow the instructions, your work may be rejected.

Step 1: Is "spoon" a correct answer? [?](#)

YES: It is correct. NO: It is incorrect.

Step 2: How many polygons are needed to locate the region that the answer "spoon" is referring to? [?](#)

Zero: No polygon is needed.
 One: Just one polygon is needed.
 More than one: Multiple polygons are needed.

Step 3: Draw one polygon to locate the region that the answer "spoon" is referring to by clicking on the image. [?](#)


1/2 answer [▶](#)

(a) User interface to ground the different answers for each visual question.

Image 1 Image 2 Image 3

Question: What is this?

Unprocessed Answer(s): plate spoon



Clear Undo Select the whole image

Please read the question and answer about the image shown on the left. Then complete the 3 steps below.
We review the results. If you do not follow the instructions, your work may be rejected.

Step 1: Is "plate spoon" a correct answer? [?](#)

YES: It is correct. NO: It is incorrect.

Step 2: How many polygons are needed to locate the region that the answer "plate spoon" is referring to? [?](#)

Zero: No polygon is needed.
 One: Just one polygon is needed.
 More than one: Multiple polygons are needed.

Step 3: Is the answer "plate spoon" already located in one of the previously drawn region? [?](#)

YES: It is already located. It is in the same location as:
 Region1: spoon

NO: It is not located. Draw polygon(s) to locate the region that the answer is referring to by clicking on the image.

[◀](#) 2/2 answer [Next Image](#)

(b) After one answer grounding was available for a visual question, the annotator could choose between selecting a previously drawn polygon as the grounding for the new answer and drawing a new polygon to ground the answer.

Figure 2: Screenshots of our annotation task interface.



Q: What letters are on the yellow part of the fire hydrant?
 Ans: albertville al



Q: What is on the fruit?
 Ans: bag



Q: What is the watermark?
 Ans: 1000 faces gregpc



Q: What is the bus number?
 Ans: 85a



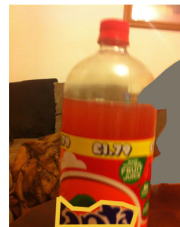
Q: Who is pulling on the other side?
 Ans: dog



Q: What color is this parking meter?
 Ans: red



Q: What type of sugar is this?
 Ans: tate lyle



Q: What is in this bottle?
 Ans: soda



Q: What's on the screen?
 Ans: starting



Q: What is this? What is this thing?
 Ans: laptop



Q: What is in this can?
 Ans: whole kernel corn



Q: What is this?
 Ans: fence



Q: what kind of coffee is this?
 Ans: dunkin donuts



Ans: dunkin donuts dunkin dark



Ans: sin city



Q: What's on this t shirt, please?
 Ans: sin city las vegas



Ans: 2 girls text sin city fabulous las vegas nevada

Figure 3: High-quality grounding annotations for visual questions where valid answers refer to different groundings. The first two rows of examples come from VQAv2 dataset and the last three rows of examples come from VizWiz-VQA dataset.



Q: What type of airplane is this?
 Ans: 747
 Ans: transport
 Ans: jet



Q: What is the wall treatment under the cabinets?
 Ans: backsplash
 Ans: brick



Q: What kind of stone is the sidewalk made of?
 Ans: rock
 Ans: cobblestone



Q: What character is wearing the red shirt?
 Ans: woman
 Ans: 1 on left



Q: What type of weather is occurring?
 Ans: raining
 Ans: rain



Q: What color is the structure?
 Ans: pink and green
 Ans: red and green
 Ans: pink



Q: What is this person doing?
 Ans: playing tennis
 Ans: tennis



Q: What type of appliance is this?
 Ans: fridge
 Ans: mini fridge
 Ans: refrigerator



Q: What time of year is it?
 Ans: winter
 Ans: christmas



Q: What color are the vegetables?
 Ans: yellow and red
 Ans: orange and red



Q: What color is the bike?
 Ans: blue
 Ans: Silver



Q: Why is this cat on the bed?
 Ans: resting
 Ans: relaxing



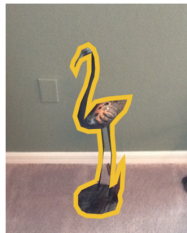
Q: Can you read the label on this bottle?
 Ans: yes
 Ans: paul mitchell
 Ans: firm style freeze
 Ans: shine super spray



Q: What is this package?
 Ans: food
 Ans: beef pot pie
 Ans: pot pie



Q: What is the picture on this t-shirt?
 Ans: fire
 Ans: flames



Q: What is this?
 Ans: bird statue
 Ans: flamingo
 Ans: bird



Q: What does that sign showing?
 Ans: pedestrian crossing
 Ans: crosswalk



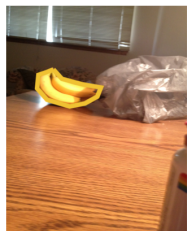
Q: Identify this product please.
 Ans: kahlua liquor
 Ans: kahlua



Q: What color, what color is it?
 Ans: tan
 Ans: beige
 Ans: khaki



Q: What color is this shirt?
 Ans: blue black
 Ans: white
 Ans: blue white striped



Q: Alright, last try on this. I think I know what it is but try to tell me please, thanks.
 Ans: banana
 Ans: bananas



Q: What is this?
 Ans: money
 Ans: \$20 bill
 Ans: 20 dollar bill



Q: What is this?
 Ans: remote control
 Ans: tv remote
 Ans: remote



Q: What color is this?
 Ans: green white
 Ans: red
 Ans: white green orange

Figure 4: High-quality grounding annotations for visual questions where all valid answers refer to the same grounding. The first two rows of examples come from VQAv2 dataset and the last two rows of examples come from VizWiz-VQA dataset.

As data collection proceeded, we leveraged a combination of automated and manual quality control steps to ensure the ongoing collection of high-quality results. For automated quality control, we calculated the mean number of times each worker selected “No” in Step 1 (contains incorrect answer), “Zero” and “More than one” in Step 2 (needs no polygon or more than one polygon) per HIT for each worker. If the mean was more than 1.25 times the mean value we observed across all workers, we randomly inspected at least ten HITs from that worker’s recent submissions. We also monitored the mean time each worker spent on each HIT. When the mean was less than 1 minute, we randomly inspected at least ten HITs from this worker’s recent submissions. Finally, we also monitored the mean of the number of points for an image (if applicable) drawn by each worker. When it was less than five points, we randomly inspected at least ten HITs from this worker’s recent submissions and provided feedback as needed. For manual quality control, we continuously reviewed random selections of submitted HITs and provided feedback, when necessary, to workers throughout the data collection process (though after the first week, we hardly noticed any issues).

Statistics for each step of filtration are shown in Table 1 to complement statistics provided in the main paper. Note that we only start crowdsourcing from a fraction of VQAv2’s training set with 9,213 overlapping with [3] and 9,000 randomly sampled.

	VizWiz	VQAv2 training	All
Original dataset	32,842	443,757	476,599
Valid Answers	9,810	164,757	174,567
Sub-questions	9,528	163,731	173,259
Crowdsourcing	9,528	[Sampled] 18,213	27,741
Incorrect Answers	7,216	8,214	15,430
No/multi polygons	6,729	5,561	12,290
75% agreement	3,442	2,383	5,825

Table 1: Number of visual questions left after each step. We filtered visual questions with less than one valid answer/answer grounding after each step if applicable.

Examples of high-quality answer grounding results are shown in Figures 3 and 4. Figure 3 shows visual questions that require *text recognition* skill tend to have different groundings for all valid answers to a visual question. Figure 4 shows visual questions that require *color recognition* tend to share the same grounding to a visual question.

II. Dataset Analysis

II.1. Incorrect Answer.

Even though we define a valid answer as at least two out of ten people agreeing on that answer, we find that 29% of answers (17,719 out of 60,526) are labeled as incorrect

Q: What color is this?
A: Light blue (incorrect)
A: Green (correct)

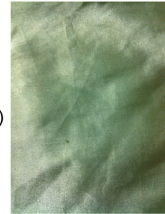


Figure 5: An example of a color-related visual question when at least two out of ten people give the same incorrect answer “light blue”.

by at least one worker; i.e., 3,309 out of 20,930 answers from VizWiz-VQA dataset and 14,410 out of 39,596 answers from VQAv2. From inspection of some of these answers, the reasons why answers are deemed incorrect are (1) regions are too small to recognize, (2) images are too low quality to recognize the content (e.g., too dark or too blurred), and (3) similar colors. For example, an image showing a green cloth while some people say it is light blue is shown in Figure 5). Examples of incorrect answers are also shown in Figure 6. Since it is hard to recognize if an answer is correct or not with the low-quality images or small groundings (e.g., the clock region is too small to tell if it is 3:30 or 12:15), we also show the correct answer and its magnified grounding for readers’ convenience.

To facilitate future work, we will share the metadata indicating which answers are “incorrect” as part of publicly-releasing our VQA-AnswerTherapy dataset. Potential use cases for identifying incorrect answers include (1) verifying provided answers in the existing VQA datasets [2, 13], which can lead to cleaner VQA datasets and (2) indicating when the model might perform even better than humans: it might be easier for the model to recognize small regions without magnifying regions and the model can also lighten, darken, or deblur images when needed. Given that a large percentage of flagged incorrect answers exist in both the VizWiz-VQA dataset [13] and the VQAv2 dataset [2], we encourage future work to explore this topic more.

II.2. No Polygon and Multiple Polygons.

Recall that when we collect the data, in step 2 we asked workers to indicate “how many polygons are needed to locate the region that the answer is referring to”. We show some visual questions when people select “no polygon is needed” and “multiple polygons are needed” in Figure 7.

II.3. Grounding Agreement.

Recall that two answer grounding annotations were collected for each unique answer per visual question from two crowdworkers.

We show a histogram of grounding alignment between two crowdworkers across the 26,682 unique image-

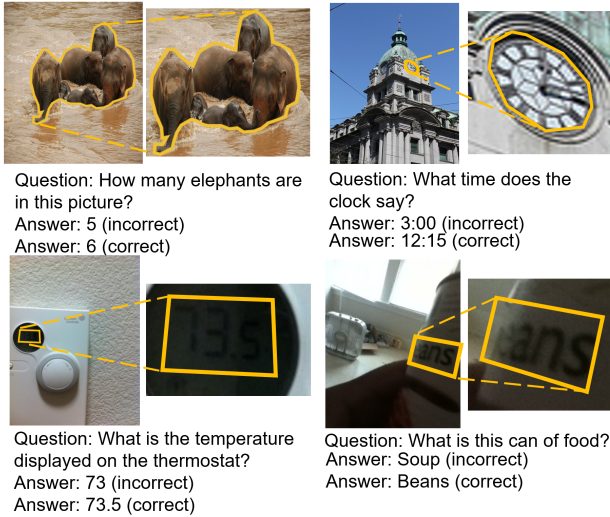


Figure 6: Examples show that when regions that lead to the answer are too small to recognize or when the image has low quality, people can answer the visual question incorrectly while achieving agreement (at least two out of ten people give the same incorrect answer). We show the correct answer and the magnified grounding for the correct answer (to the right of the original image) for readers’ convenience because, without magnifying, some regions that lead to the answers are too small/too low quality to tell whether the provided answers are correct or not.

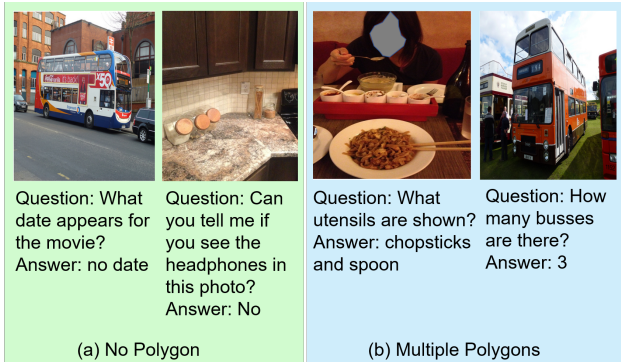


Figure 7: Visual questions that (a) have no answer grounding (i.e., need no polygons) and (b) need more than one polygon for the answer grounding.

question-answer triples in Figure 8. The majority (53%, 14,262 out of 26,682) of the IoU scores are between 0.75 and 1.0, ~20% (5,101 out of 26,682) between 0.75 and 0.5, ~10% (2,865 out of 26,682) between 0.5 and 0.25, and 17% (4,453 out of 26,682) lie between 0.25 and 0. We attribute grounding misalignments largely to the grounding being ambiguous, as exemplified in the first row of Figure 9, and redundant information in the image where different regions can independently indicate the same answer, as ex-

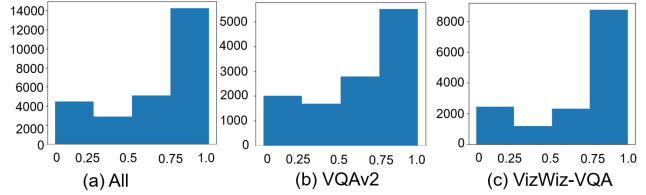


Figure 8: Histogram of IoU scores indicating similarity between each pair of answer groundings per visual question. The majority have a high agreement, in the range between 0.8 and 1.0.

emplified in the second row of Figure 9.



Figure 9: Examples of low alignment between two workers’ annotations because of ambiguous or redundant information where different regions can independently indicate the same answer.

The grounding differences from different workers highlighted a few questions that we leave for future work: (1) When grounding an answer, should we ground all the information (both the explicit information and the implicit information) that leads to the answer, or just explicit information?, (2) Should we ground all the information or just part of information (e.g., many regions independently lead to the same answer and we just ground the most obvious one) if part of the information is already sufficient?, (3) When workers draw regions that are highly aligned with

IoU	All		VQAv2		VizWiz-VQA	
	Single	Mult	Single	Mult	Single	Mult
0.7	5027	798	2245	138	2782	660
0.75	4992	833	2243	140	2749	693
0.8	4957	868	2238	145	2719	723
0.85	4932	893	2235	148	2697	745
0.9	4909	916	2228	155	2681	761
0.95	4896	929	2225	158	2671	771
1	4889	936	2223	160	2666	776

Table 2: Number of VQAs with a single grounding and multiple (Mult) groundings under different IoU thresholds.

each other, which grounding should we select?

II.4. Reconciling Redundant Annotations.

As mentioned in the main paper, during the annotation process, we allow workers to select each answer if the answer has been located in one of the previously drawn regions (See Figure 1 Step 3 - Option (a) and Figure 2 Step 3). Then we selected the larger grounding from two groundings if the two groundings’ alignment is larger than 0.75. We observe that frequently (93%, i.e., for 5,459 out of 5,825 visual questions), the selected answer grounding for different answers to one visual question are from the same worker (recall though that the annotations across different visual questions still can come from different workers). For VizWiz-VQA, 3153 visual questions each have all answer groundings coming from the same worker and 289 from different workers. For VQAv2, 2306 visual questions each have all answer groundings coming from the same worker and 77 from different workers. These facts highlight that a visual question’s different answer groundings can all be identical and so have an IoU = 1.0.

We decide whether the answer groundings are based on the same regions by calculating IoU scores for every possible answer grounding pair per visual question and checking if all of the grounding answer pairs have an IoU score larger than 0.9. If their overlap is larger than 0.9, we believe this visual question has the same grounding for all answers. We chose an IoU threshold less than 1.0 to accommodate the 7% of visual questions where different answer groundings for the same visual question came from different workers. We also report in Table 2 the number of visual questions identified as having a single versus multiple groundings when using different IoU thresholds between 0.9 and 1.0. The results show similar outcomes when using different thresholds.

II.5. Four Grounding Relationships.

We visualize four kinds of relationships, i.e., disjoint, equal, contained, and intersected, between every possible

answer grounding pair in Figure 10. These exemplify that visual questions needing *object recognition* tend to have disjoint or contained relationships, visual questions needing *text recognition* tend to have intersected relationships, and visual questions needing *color recognition* tend to have an equal relationship.



Figure 10: For each visual question, we flag which relationship types arise between every possible answer grounding pair from the following options: disjoint, equal, contained, and intersected.

II.6. Most Common Answers.

Due to space constraints, we provide the analysis of the most common answers that co-occur with a single grounding here. We obtain the most common answers following a similar process as used to obtain the most common questions in the main paper. The top five common answers

for the VQA-AnswerTherapy dataset that co-occur with a single grounding are ‘white’, ‘phone’, ‘blue’, ‘black’, and ‘brown’. The top five common answers for VQAv2 are ‘white’, ‘brown’, ‘black’, ‘gray’, and ‘blue’. The top five common answers for VizWiz-VQA are ‘phone’, ‘grey’, ‘blue’, ‘remote’, and ‘remote control’. We show the Word-Cloud for the common answers that lead to the same answer grounding for VQA-AnswerTherapy as well as for the VizWiz-VQA and VQAv2 datasets independently in Figure 11. These findings reinforce our conclusion in the main paper that visual questions requiring object or color recognition skills tend to share the same groundings.

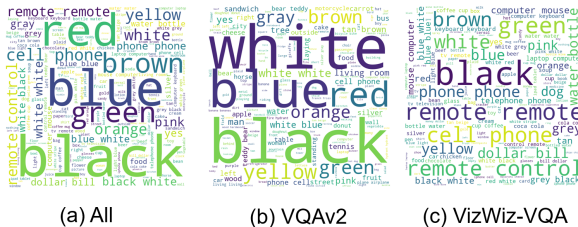


Figure 11: Most common answers for visual questions that have the same groundings for all unique answers.

III. Algorithm Benchmarking

Experimental Details for Single Answer Grounding Challenge. We used an AdamW optimizer with a learning rate of 0.00005 and fine-tuned ViLT on the VizWiz-VQA and VQAv2 datasets for 20 epochs.

For mPLUG-Owl, we did preliminary testing with four different prompts and selected the best one:

“The following is a conversation between a curious human and AI assistant. The assistant only replies “YES” or “NO” to the user’s questions.

Human: <image>

Human: What are all plausible answers to the question <INSERT QUESTION VARIABLE>?

Human: Do all plausible answers to the questions <INSERT QUESTION VARIABLE> indicate the same visual content in this image? Reply “YES” or “NO”.

AI: ”.

The responses from mPLUG-Owl were typically either “yes” or “no” followed by a reason, (even though the model was instructed not to respond with reason). We converted the first three characters of each response to lowercase and then compared them to the ground truth to see if there is a match. If the response is anything other than “yes” or “no,” we disregard it as it cannot be reflected in precision or recall. There are 10 out of 496 samples that don’t have “yes” or “no” as their first three characters in the VQAv2 dataset, and there are 16 such instances out of 889 samples in the VizWiz-VQA dataset.

Models	All	VQAv2	VizWiz-VQA
SeqTR (I+Q+A)	66.26	64.34	67.33
SeqTR (I+Q)	61.62	58.30	63.47
SeqTR (I+A)	62.91	57.97	65.67
SEEM (I+Q+A)	53.15	50.13	54.84
SEEM (I+Q)	44.65	44.39	44.80
SEEM (I+A)	51.64	46.50	54.51
UNINEXT (I+Q+A)	48.39	42.28	59.34
UNINEXT (I+Q)	45.88	40.76	55.06
UNINEXT (I+A)	47.45	41.26	58.55

Table 3: mIoU-PQ Performance of three models on our dataset.

Experimental Details for Answer(s) Grounding Challenge. For SeqTR model, we used the pre-trained RefCOCOg weights from the SeqTR author’s repository (<https://github.com/sean-zhuh/SeqTR>) and fine-tuned it for 5 epochs following the author’s guidelines.

For the UNINEXT model, we used UNINEXT’s second stage pre-trained weights, which were top-performing for COCO detection and segmentation (verified by author). Of note, UNINEXT is also pretrained on RefCOCO and so was exposed to the COCO images utilized in our dataset. For SEEM model, we used the SEEM-FOCAL-V1 checkpoint from author’s repository (<https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once>). SEEM is also pre-trained on RefCOCO and COCO2017 and so was also exposed to the COCO images utilized in our dataset.

III.1. Performance of Three Models for Answer(s) Grounding Challenge with IoU-PQ Metric

We show the IoU-PQ performance in Table 3, the results and observations are highly aligned with the mIoU metric reported in our main paper.

III.2. Answer(s) Grounding: Qualitative Results for Model Benchmarking.

We provide additional qualitative results here for the Answer(s) Grounding task for the top-performing set-up where we feed models the image, question, and answer. . Examples are provided in Figures 12, 13, 14, and 15.

Figures 12 and 13 show visual questions with different answers that lead to the **same groundings**. Overall, we observe that models can predict well for this case, particularly when grounding a single dominant object on a relatively simple background. However, if the picture is captured from an unusual perspective or shows multiple objects (e.g., 12 “Is the truck pulling something”), models can fail. We also observe that though the answers are referring to the same region, the model’s predictions for different answers

sometimes can differ. This is exemplified in Figure 12's column 1 for SEEM(I+Q+A) ("What color is the ball") and column 2 for SEEM (I+Q+A) ("What is on other side of river").

Figures 15 and 14 show the qualitative results for the models tested on visual questions with different answers that lead to **multiple groundings**. Though different answers can refer to different regions, the model's predictions for different answers are sometimes the same. The model might perform better when identifying common objects when the camera directly faces the object (e.g., shown in Figure 14's column 1 ("What is sitting on the table?")) and worse when the content of interest is captured from other perspectives (e.g., Figure 15's column 2 ("What's that?")). The model also fails to distinguish regions for different text related answers, as exemplified in Figure 14's column 3 ("What brand logos are visible in this image?") and Figure 15's column 3 ("What kind of coffee is this?") and column 4 ("What does this say?").

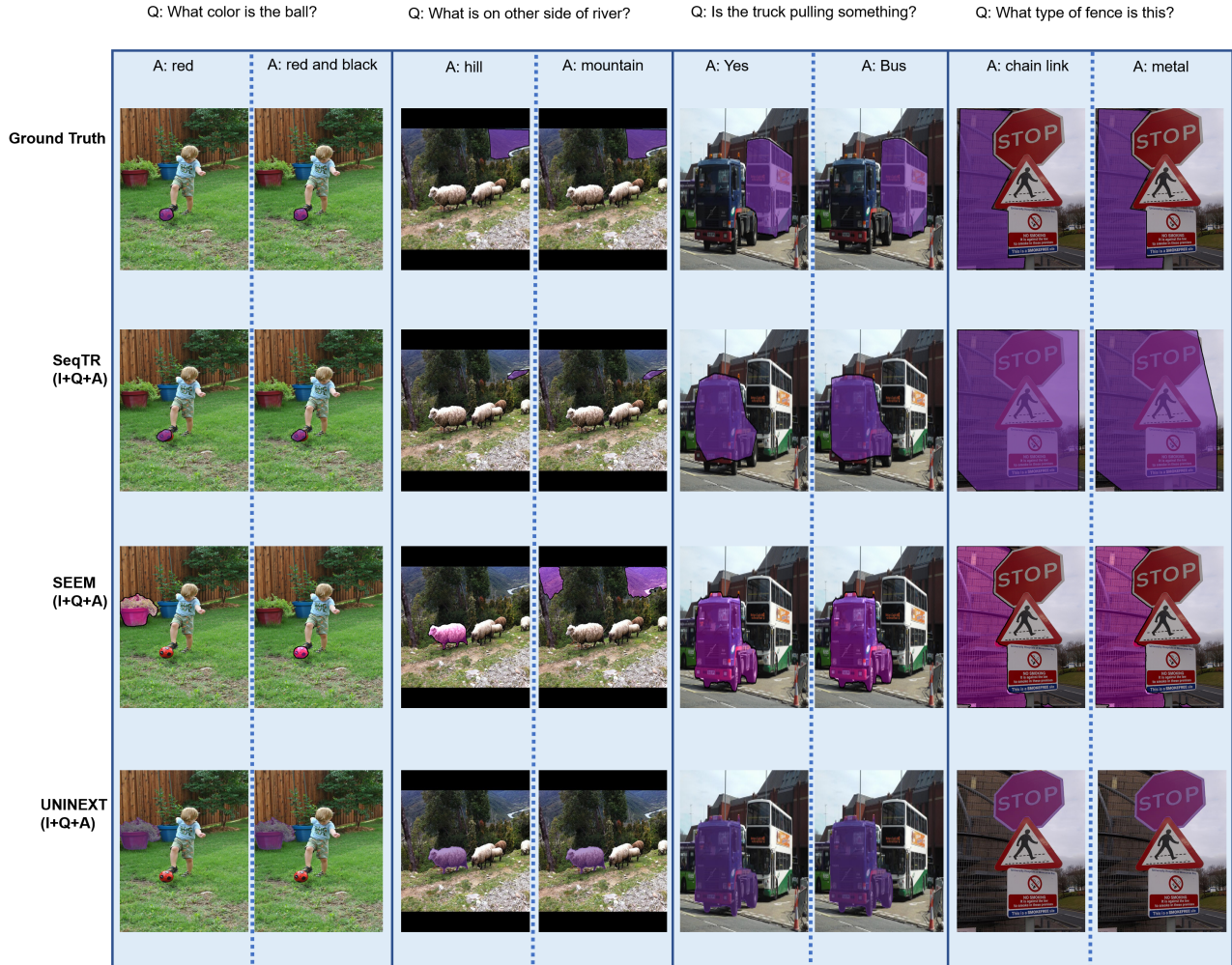


Figure 12: Qualitative results for models tested on visual questions with different answers leading to **same groundings**. Image sources are VQA_{v2} datasets (in the blue background). For each visual question, the first row shows the ground truth grounding area, the second, third, and fourth row show groundings generated by different models. Each column shows the grounding for an answer.

	Q: What kind of a bottle is this?		Q: Which color is this shirt?		Q: What is this?		Q: The expiration date?	
	A: glass	A: jar	A: tan	A: white	A: phone	A: cell phone	A: January 27 2013	A: Jan 27 2013
Ground Truth								
SeqTR (I+Q+A)								
SEEM (I+Q+A)								
UNINEXT (I+Q+A)								

Figure 13: Qualitative results for models tested on visual questions with different answers leading to **same groundings**. Image sources are VizWiz-VQA datasets (in the yellow background). For each visual question, the first row shows the ground truth grounding area, the second, third, and fourth row show groundings generated by different models. Each column shows the grounding for an answer.

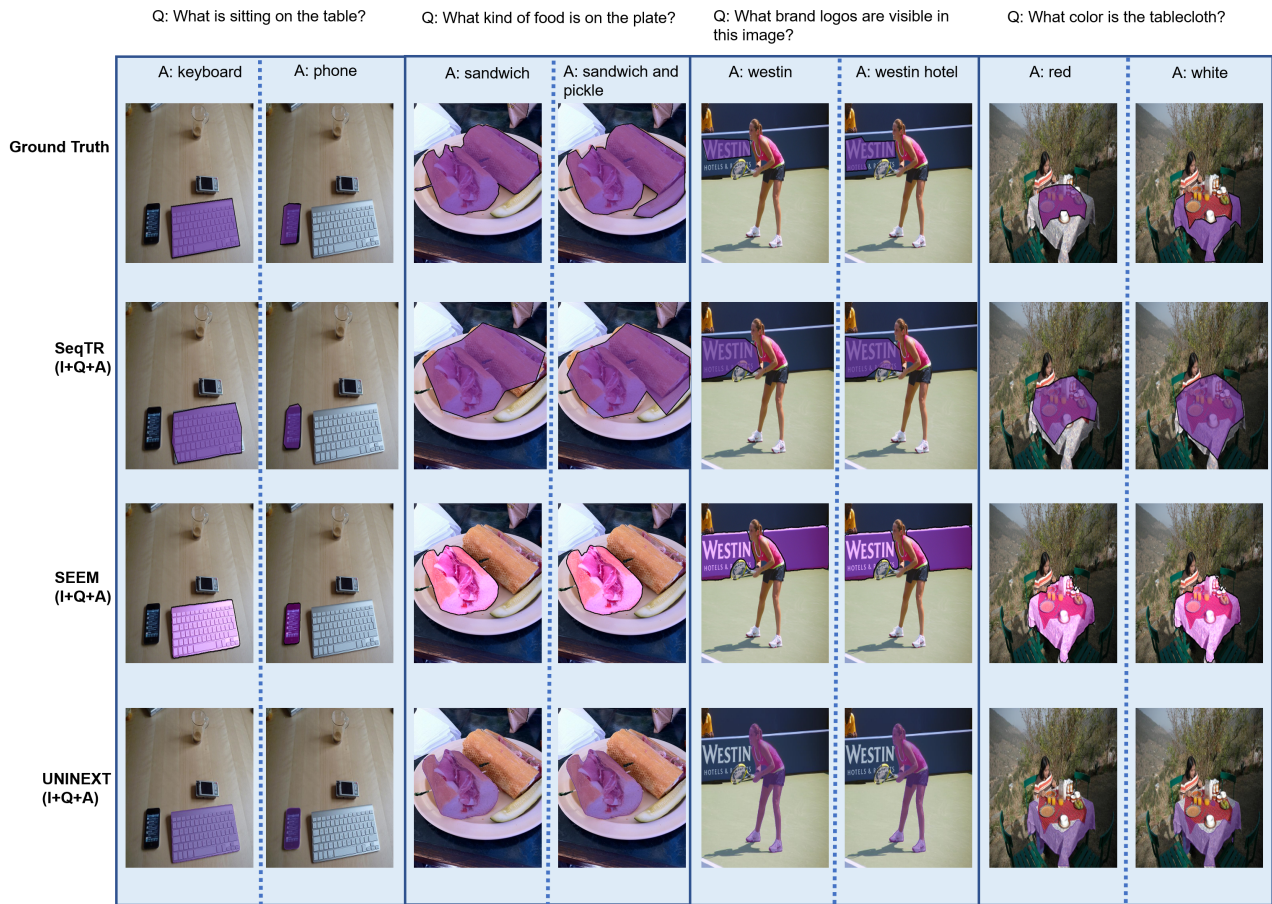


Figure 14: Qualitative results for models tested on visual questions with different answers leading to **different groundings**. Image sources are VQA_{v2} datasets (in the blue background). For each visual question, the first row shows the ground truth grounding area, and the rest of the rows show the models' predicted area. Each column shows the grounding for an answer.

	Q: What is it?		Q: What's that?		Q: What kind of coffee is this?		Q: What does this say?	
	A: dog	A: wheelbarrow	A: desk	A: office	A:colombian	A: van houtte colombian	A: Insert money card here	A: Insert moneycard here for card balance to add value
Ground Truth								
SeqTR (I+Q+A)								
SEEM (I+Q+A)								
UNINEXT (I+Q+A)								

Figure 15: Qualitative results for models tested on visual questions with different answers leading to **multiple groundings**. Image sources are VizWiz-VQA (in the yellow background). For each visual question, the first row shows the ground truth grounding area, and the rest of the rows show the models' predicted area. Each column shows the grounding for an answer.