Grounding Answers for Visual Questions Asked by Visually Impaired People: Supplementary Material

Chongyan Chen[1], Samreen Anjum[2], Danna Gurari[1,2]
[1] University of Texas at Austin [2] University of Colorado Boulder

This document supplements the main paper with more information about:

1. Method for filtering visual questions (supplements Section 3.1)
2. Dataset collection (supplements Section 3.1)
    - Method for hiring expert crowdworkers (supplements Section 3.1)
    - Screenshot of our annotation task interface (supplements Section 3.1)
    - Method for reviewing work from crowdworkers (supplements Section 3.1)
3. Dataset analysis (supplements Section 3.2)
4. Qualitative results for model benchmarking (supplements Section 4)

I. Method for filtering visual questions

Our aim was to ensure our dataset focused on visual questions for which answers could be unambiguously grounded to a single region. To do so, we performed five rounds of filtering visual questions from the initial VizWiz-VQA dataset, that we decribe below.

First, we applied a filter to *remove all questions which were unanswerable*. This occurs regularly in the VizWiz-VQA dataset because the photographers could not verify the content in their images due to being blind. We removed all visual questions which were labelled as "unanswerable" in the provided "answer_type" metadata.

We also removed visual *questions for which at least half of the crowd did not agree on the same answer*. We used a stringent string matching approach to detect if at least 5 out of the 10 answers per visual question are identical.

Another filter we applied is to remove all so-called *"questions" that actually embed multiple questions*. An example is "What is this and is it to be put in the microwave? Or does it even say?" We did not simply filter visual questions with more than two question marks because we observed that users often ask refinements to their questions or ask the same question several times in different ways. Examples include "What color are these jeans? Pink or gold?" and "Can I wear these two pieces together? Do they match?". From pilot testing, we observed an effective automated mechanism is to remove visual questions that have more than five words while containing the word "and" (i.e., 901 visual questions). We also process those visual questions that contained repetition of a single question, e.g., "what is this? what is this?". We trim these down to a single question to remove redundancy. We then had two expert crowdworkers review each candidate visual question to identify whether each question actually asked more than one question. We filtered a visual question if at least one person flagged it as containing more than one question. Of the 11,085 visual questions that were reviewed, 0.6% (66) were tagged as containing more than one question.

We also filtered visual *questions for which there is ambiguity where the answer is located in an image*, meaning answers referred to more than one image region. We conducted preliminary analysis to understand the prevalence of this case. From 200 random visual questions from the VizWiz training dataset, we found only two contained this issue. From further analysis, we found that this issue often appears when the questions contains plural nouns[1] or when they contain phrases such as "how many". While the occurrences were rare, we still had two expert crowdworkers review each VQA to identify whether more than one polygon is needed to locate the region to which the answer is referring. A VQA was removed if at least one person flagged it as needing more than one polygon. Of 11,019 visual questions that were reviewed, 1.66% (183) were marked as requiring more than one polygon to locate the answer grounding region. From inspection of some of these visual questions, possible reasons for why an answer refers to more than one region are that multiple regions are suitable for arriving at the same answer (Figure 1a) and that the answer actually embeds multiple answers that align with distinct visual content (Figure 1b).

Finally, we filtered visual *questions that were answer-*

[1] We used the NLTK package to detect parts of speech, including plural nouns.

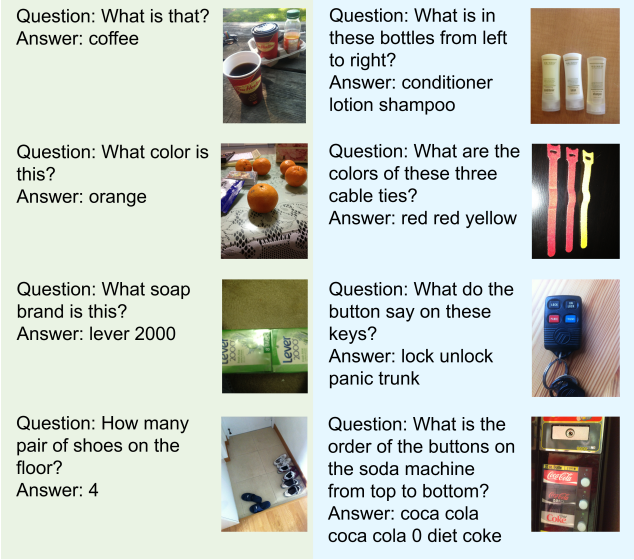

Figure 1. Answer groundings that refer to multiple regions.

able but could not be grounded. For example, often this occurs for questions that lead to the answer "No", such as "Is there a key?". To do so, we had two expert crowdworkers review each VQA triplet to identify whether the answer is not shown in the image. A VQA was removed if at least one person flagged it as not present.

II. Dataset Collection

Method for hiring expert crowdworkers

We only accepted candidates who previously had completed at least 500 Human Intelligence Tasks (HITs) with over a 95% acceptance rate and were from the United States. The latter requirement gave us some confidence that the workers had English proficiency.

Next, we required the crowdworkers to pass a qualification test which included challenging grounding tasks covered in our instructions. In this test, each worker is asked to annotate 10 QA pairs for which the ground truth (GT) annotations were manually drawn by us. The grounding annotation by the worker is deemed correct if the region has more than a 70% IoU score with the GT region. The workers had to annotate all of the 10 QA pairs correctly to pass the qualification test. The user interface blocked a worker from moving to the next of the 10 tasks until the generated annotation sufficiently matched our pre-annotated ground truth. Completing this qualification task ensured a worker understood the task and how to handle challenging annotation scenarios.

All workers who passed the qualification test were eligible to complete 20 grounding tasks. In total, 27 workers completed these tasks. We reviewed all annotations from them and hired nine workers who consistently generated high quality results. We limited our number of workers because we prioritized high quality annotations more than the efficiency from having more workers; i.e., it is easier to track the performance of fewer workers.

Screenshot of our annotation task interface

We show the crowdsourcing task that we created to collect grounding annotations, including a screen shot of the instructions in Figure 2 and the annotation task in Figure 3. The link to this code is available at https://github.com/CCYChongyanChen/VizWizVQAGroundingCrowdSourcing.

Method for reviewing work from crowdworkers

The nine workers hired to create all answer groundings were given our contact information so they could contact us with any questions and we gave them the link to a live document where we frequently added our feedback to their questions about tricky examples. As they submitted their work, we leveraged automated quality control checks and manually inspected random samples of their results to ensure the annotation quality remained high. For the automated checks, we used the following rules to help us identify quality issues. For each HIT, we recorded the total number of times a worker answered 'Yes' in Step 1 (contains multiple questions), 'Yes' in Step 2 (contains multiple regions), and 'cannot draw' in Step 3 (rather than drawing a segmentation). If the total number for any of these steps was more than 3, we reviewed all results in the HIT. This is because, our pre-processing steps (described in Section I) meant that most of the QA pairs should contain one question and our preliminary analysis indicated that we seldom observed either that there are multiple regions or that the answer is not present in the image. We also monitored the time the worker spent on each HIT. If it was less than 30 seconds, we inspected the results.[2] Finally, we also calculated the number of points used to draw the polygon. If a worker drew less than 5 points for an answer grounding for more than two visual questions in a single HIT, we manually inspected all results for the HIT. Examples of high quality answer grounding results are shown in Figure 4.

[2]In pilot studies, we found that initially crowdworkers took an average of 4.38 mins to finish a HIT but this time dropped to 2.8 mins as workers became familiar with the task.

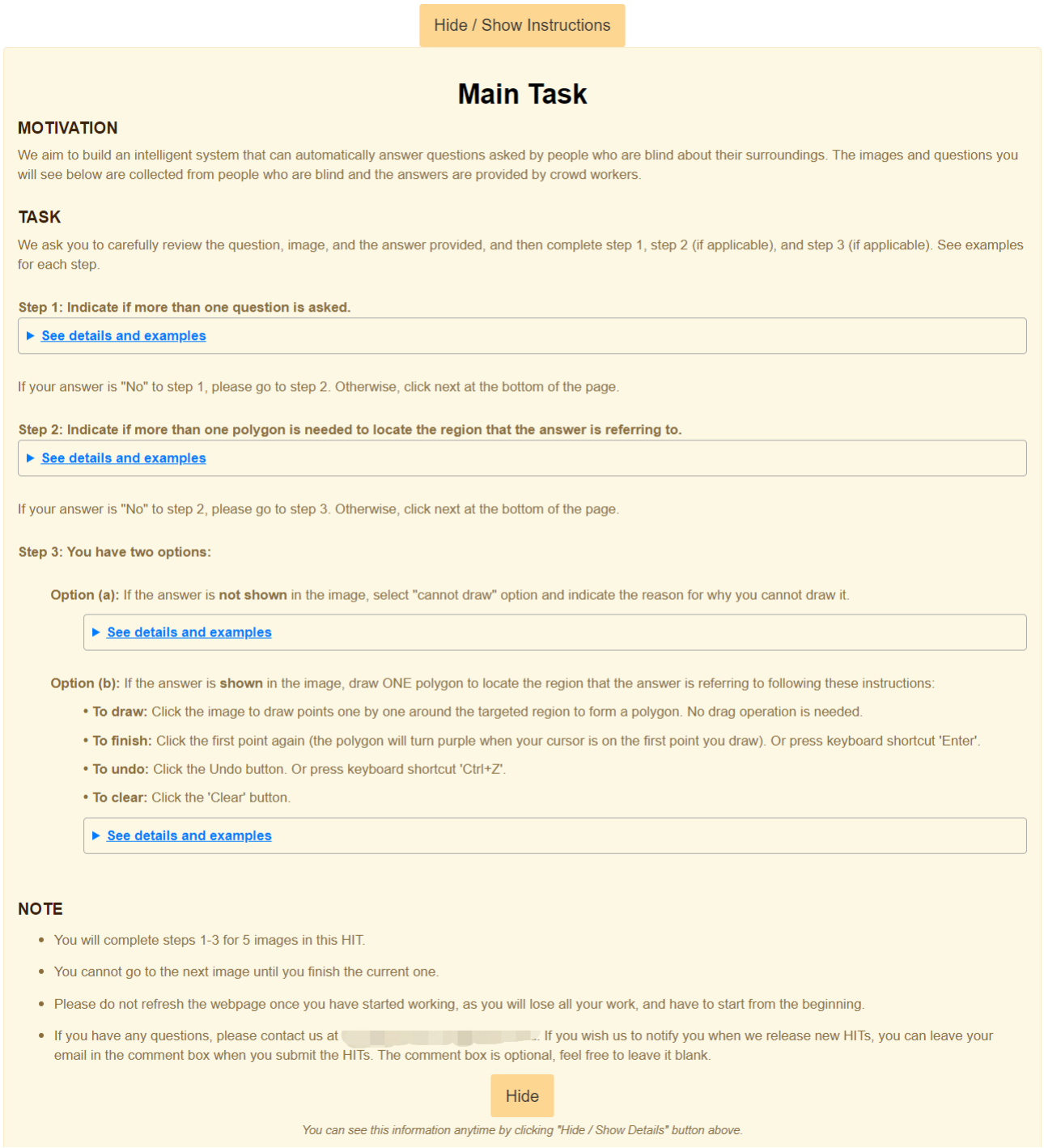

Figure 2. Instructions for our annotation task interface.

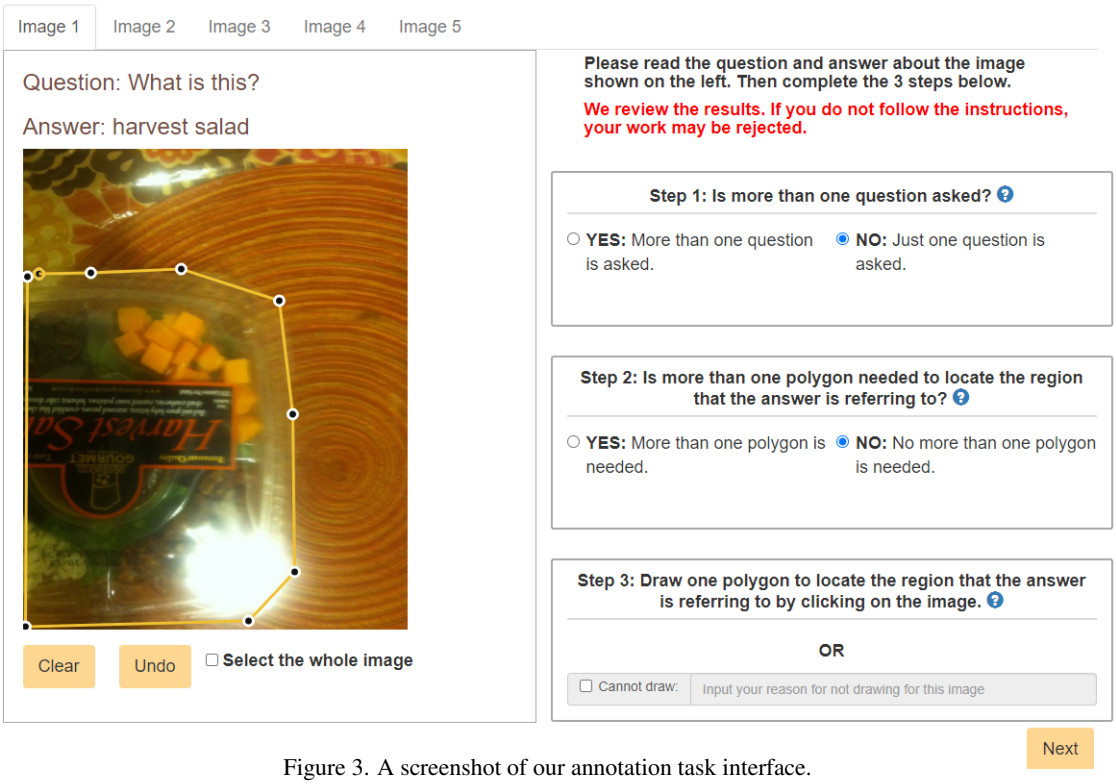

Figure 3. A screenshot of our annotation task interface.

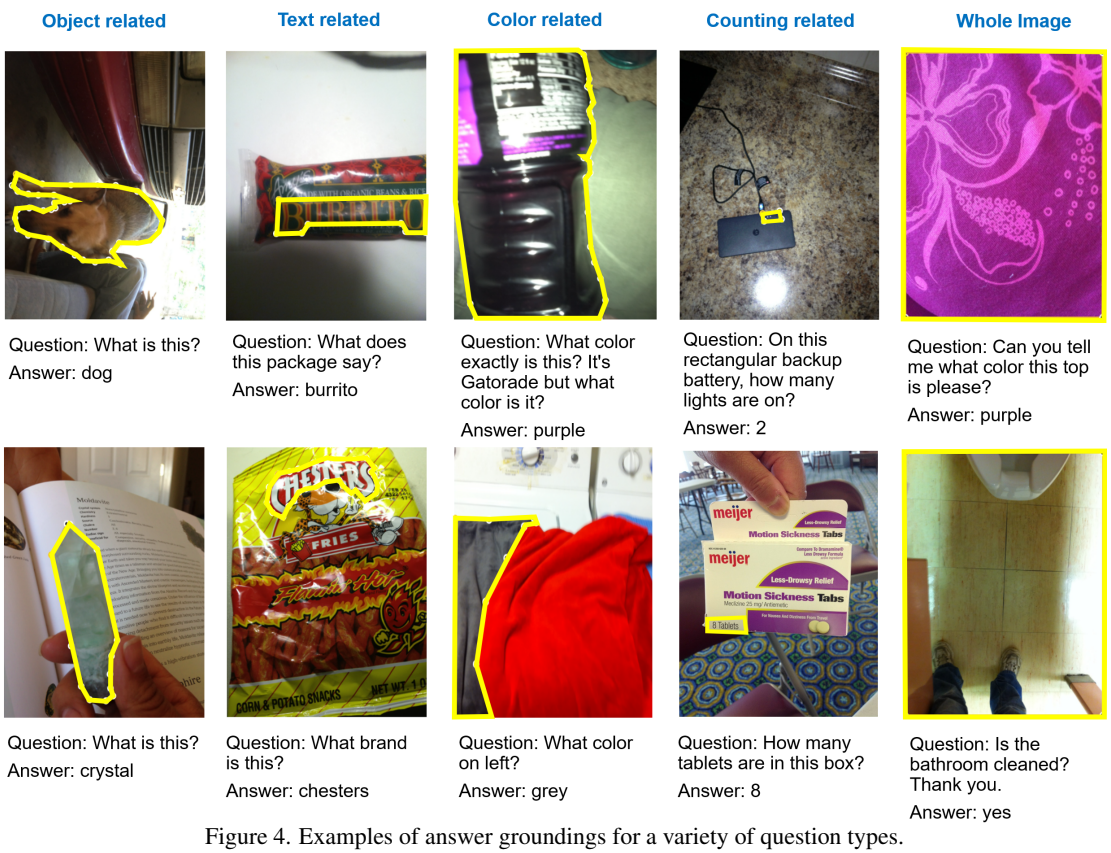

Figure 4. Examples of answer groundings for a variety of question types.

III. Dataset Analysis

Reasons why crowdworkers selected "cannot draw". The crowdworkers explained why they indicated the answer cannot be localized in the image (i.e., by selecting "cannot draw" in the task user interface) in a free-text box. The top 10 reasons are 'nothing to draw a polygon' (106 times), 'incomplete text' (28 times), 'wrong answer' (25 times), 'the image is too blurry/blurry' (26 times), 'no clue(s)' (20 times), 'subjective' (9 times), 'the answer is no so therefore can't draw something not there' (7 times), 'nothing on screen' (4 times), 'answer is not shown in the image and the answer can be answered without the image' (4 times), and 'nothing found' (4 times). Examples of these flagged visual questions are shown in Figure 5.

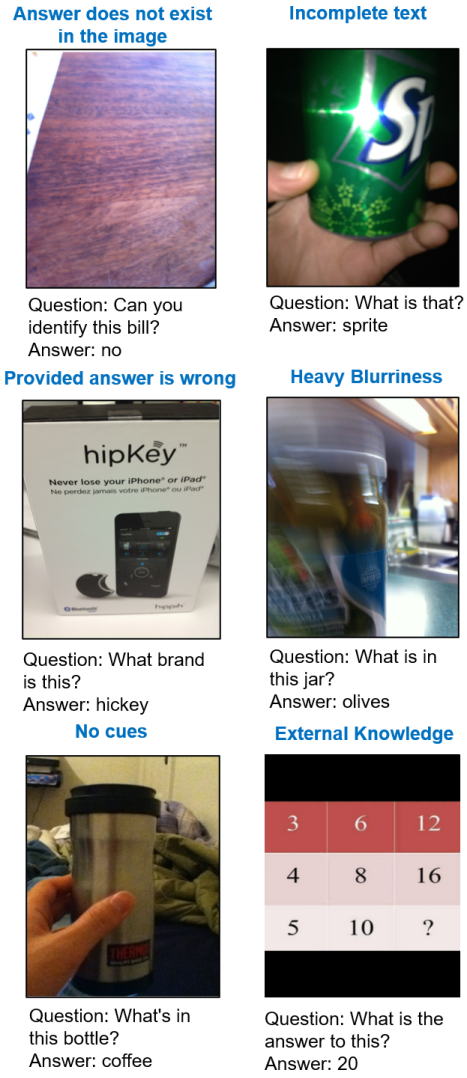

Figure 5. Examples of visual questions for which workers indicated the answer cannot be grounded in the image paired with a more general reason why (shown in blue text).

Annotation agreement. Recall that two answer groundings annotations were collected per visual question from two crowdworkers. A histogram showing the IoU scores between each pair of annotations per visual question across the 9,998 visual questions is shown in Figure 6. The majority (around 60%) of the IoU scores are between 0.8 and 1.0, around 20% lie between 0.6 and 0.8, and slightly more than 10% lie between 0 and 0.2. This shows that, typically, there is high annotation agreement. We attribute low scores in part to the IoU being a poor metric when accounting for smaller regions. An example is shown in Figure 7, where the IoU score is ∼58% despite that both annotations are visually similar and correct. We selected as ground truth the larger region from the two groundings since often the smaller one is contained in the larger one.

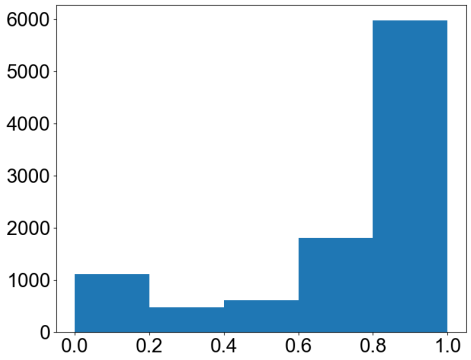

Figure 6. Histogram of IoU scores indicating similarity between each pair of answer groundings per visual question. The majority have high agreement, in the range between 0.8 and 1.0.

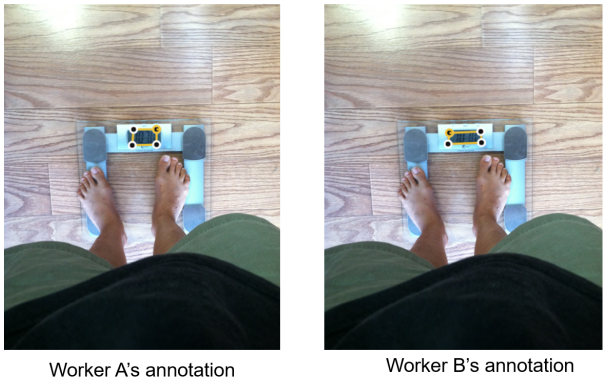

Figure 7. Example of two answer groundings for a small area. The IoU score between these two annotations is 58%. This exemplifies the tendency for IoU scores to be low for small regions when groundings appear similar.

Whole Image The visual answer is labelled as referring the whole image for 0.9% (903) of visual questions. Often, workers selected "Whole Image" when the question related to color, the camera is set too close to the object, or the questioner asks about the general description of the scene. Examples are shown in the last column of Figure 4.

Location of answer grounding. Expanding on Table 1 in the main paper, we visualize the center of mass for all answer groundings in the different datasets. Results are shown in Figure 8. As shown, our new dataset clusters as a circle and has a smaller range of values than the other datasets.

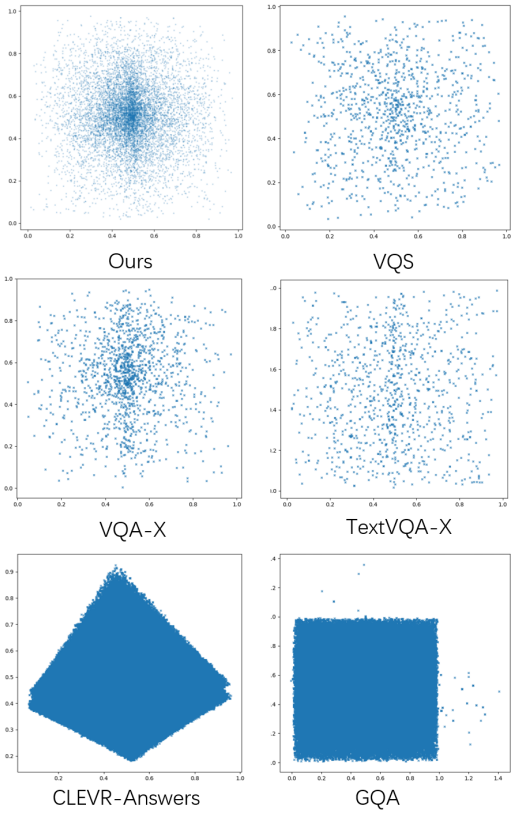

Figure 8. Center of the mass of answer groundings in the VizWiz-VQA-Grounding, VQS, VQA-X and the TextVQA-X datasets.

Grounding of most common answers. Expanding on the analysis in the main paper, we report statistics about the most common answers in Table 1. Intuitively, we observe that objects with rectangular shapes need fewer points, e.g., keyboard, laptop, and shampoo. In contrast, objects with complex boundary require more points, e.g., dog, hand, chair, and cat. Less intuitively, the mean of the grounding size for the answer 'yes' is smaller than those for which the answer is 'no'. We suspect groundings for 'no' answers more often refer to the whole image while 'yes' primarily focus on specific regions, as exemplified in Figure 9.

Answer	Size (%)	Points	Images
yes	89,264 (45%)	13	525
no	112,648 (56.3%)	10	240
keyboard	125,786 (62.9%)	7	118
dog	63,598 (31.8%)	40	85
laptop	122,471 (61.2%)	9	66
pepsi	55,055 (27.5%)	12	48
coca cola	47,644 (23.8%)	11	46
orange	82,201 (41.4%)	14	42
corn	57,947 (29.0%)	11	37
green beans	56,246 (28.1%)	10	34
pen	16,731 (8.4%)	17	33
lotion	42,835 (21.4%)	16	32
cat	54,714 (27.4%)	34	30
water bottle	61,062 (30.5%)	24	28
phone	98,399 (49.2%)	15	28
soup	44,875 (22.4%)	11	28
Shampoo	35,190 (17.6%)	8	26
hand sanitizer	68,540 (34.3%)	23	26
hand	86,867 (43.4%)	39	26
chair	81,435 (40.7%)	37	26
remote	63,090 (31.5%)	17	26

Table 1. Mean value of properties describing the regions for most common answers (excluded color-related answer, as it is reported in Table 2.

Figure 9. Example illustrating the trend that the 'yes' answer groundings are typically smaller than the 'no' answer groundings.

We show for the most common answers examples of the grounded area as well as the average images in Figure 10. For the same language answer, visual groundings can be diverse. For example, the dog has different breeds, colors, postures, and locations. The dog can be partially visible and under different illumination. Also, the answer "dog" can refer to an animal or a picture of a dog. A more blurry/grey average image is indicative of a greater diversity of images for the answer.

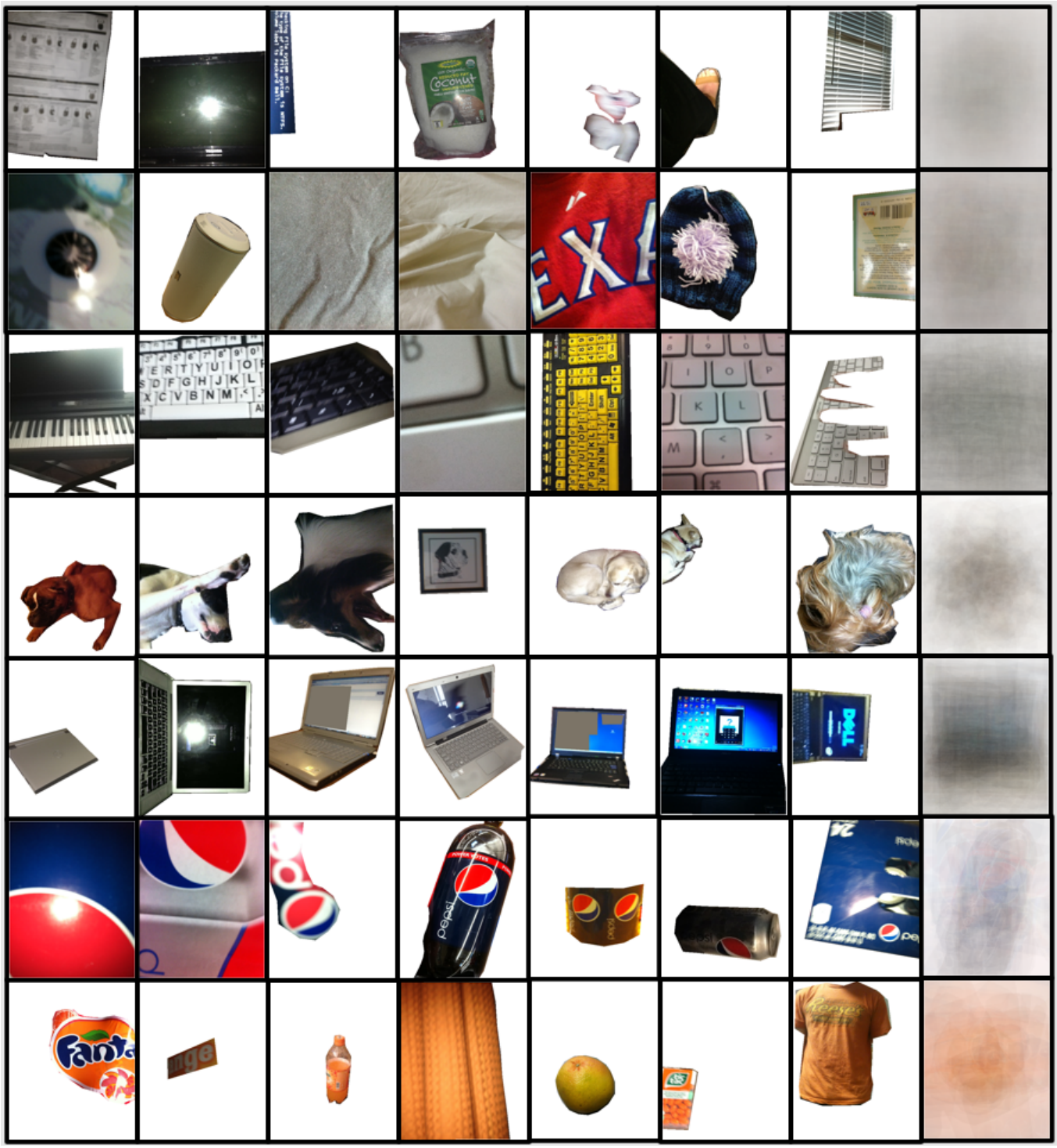

Figure 10. Examples of answer groundings in our VizWiz-VQA-Grounding dataset. Shown are answering groundings for common answers (i.e., yes, no, keyboard, dog, laptop, pepsi, orange) as well as the average image across all groundings for each answer.

Dataset comparison. We summarize how the characteristics of the eight related answer grounding datasets, discussed in the related work section of the main paper, relate and differ to our dataset in Table 2. Our summary indicates the visual annotation type, number of images with visual annotation, and source VQA dataset.

Dataset	Visual Annotation Type	# Images (× Annotations per Image)	VQA dataset
Visual7W (2016) [15]	Bounding box	47,300	COCO [9]
VQA-HAT (2017) [3]	Human Att. (deblur image)	59,849	VQA v1
VQS (2017) [4]	Segmentation+bounding box	37,868	COCO
VQA-X (2018) [7]	Segmentation	6,000 (× 1)	COCO, VQAv2
GQA (2019)	Bounding box	355,530 (×1)	GQA (2019) [6]
AiR (2020) [1]	Human Att. (eye-tracking)	987 (× 20)	GQA [6]
Text VQA-X (2021) [10]	Segmentation (brush)	11,681(× 1)	TextVQA [11]
CLEVR-Ans (2021) [13]	Bounding box	445,268 (×1)	CLEVR
Ours	Segmentation	9,998 (× 2)	VizWiz-VQA [5]

Table 2. Comparison between existing VQA answer grounding datasets and our dataset.

IV. Algorithm

mAP@IoU results. The performance for each model on the VizWiz-VQA-Grounding test split with respect to mAP@IoU score is shown in Table 3. Overall, the low mAP scores reinforce our findings in the main paper that the models perform poorly on our new dataset and that the best indicator of better answer groundings is that models were pre-trained on the VizWiz-VQA dataset (Sec 4).

Model (Pretrained)	mAP25	mAP50	mAP75	mAP
LXMERT (VizWiz)	14.21%	1.99%	0.02%	0.49%
OSCAR (VQA-v2)	7.10%	0.15%	0.00%	0.03%
MAC-Caps (GQA)	3.94%	0.09%	0.00%	0.02%
MAC-Caps (CLEVR)	6.38%	0.12%	0.00%	0.01%
MAC-Caps (VQA-v2)	9.32%	0.45%	0.01%	0.08%
MAC-Caps (VizWiz)	**22.32%**	**4.09%**	**0.17%**	**0.96%**

Table 3. Performance of six models when evaluated on the VizWiz-VQA-Grounding test set: two state-of-art VQA models (LXMERT [12] and OSCAR [8]) and four variants of the state-of-art VQA model for answer grounding (MAC-Caps [13]) with respect to mAP@IoU. Results are provided based on the COCO evaluation protocol of using different IoU thresholds, from 0.25 to 0.75, and averaging AP values with IoU threshold ranges from 0.5 to 0.95 with a step size of 0.05

Extraction of attention maps: baseline models. As discussed in the main paper, we extract attention maps for the two VQA models: LXMERT and OSCAR. For LXMERT, the output consists of four different attentions: self-vision attention, self-language attention, image-guided question attention, and question-guided image attention. Following the authors' recommendation, we picked the question-based image attention. For each visual question, the model predicts 12 attention heads, where each head was of size 20x36. We first average the attention maps for each head by length, and then over all the heads to obtain the final attention map. For OSCAR, we extract attention weights from the last layer. It has 114x114 for 16 heads; For the 114 sequence length, the first 64 dimensions are the language attention, while the later 50 dimensions are image attentions. We uses self-vision attention.

Naive baseline. As a naive baseline, we treated predicting the whole image as the grounded area. This baseline receives an average IoU score of 33%.

Analysis With Respect to Image Quality. As mentioned in the main paper, we report here our fine-grained analysis to assess each model's ability to accurately locate the answer groundings based on the image quality issues defined in [2]: poor framing, blurry, too dark, too bright, obfuscations, and improper rotations. Results are shown in Figure 11. We observe that images with obscured image quality issue are the most challenging to ground and images with the rotation issue are the second most challenging to ground.

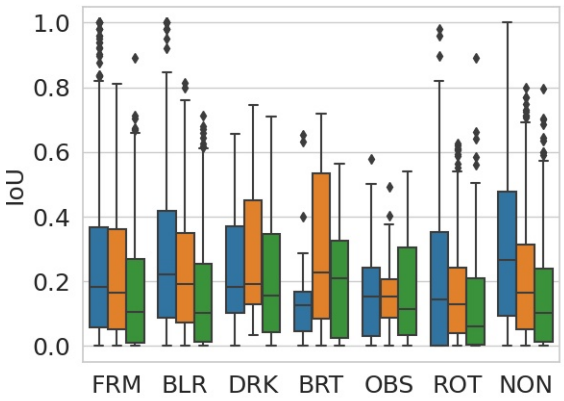

Figure 11. Comparison of MAC-Caps (pretrained on VizWiz), LXMERT, and OSCAR's performance on visual questions for images with different quality issues.

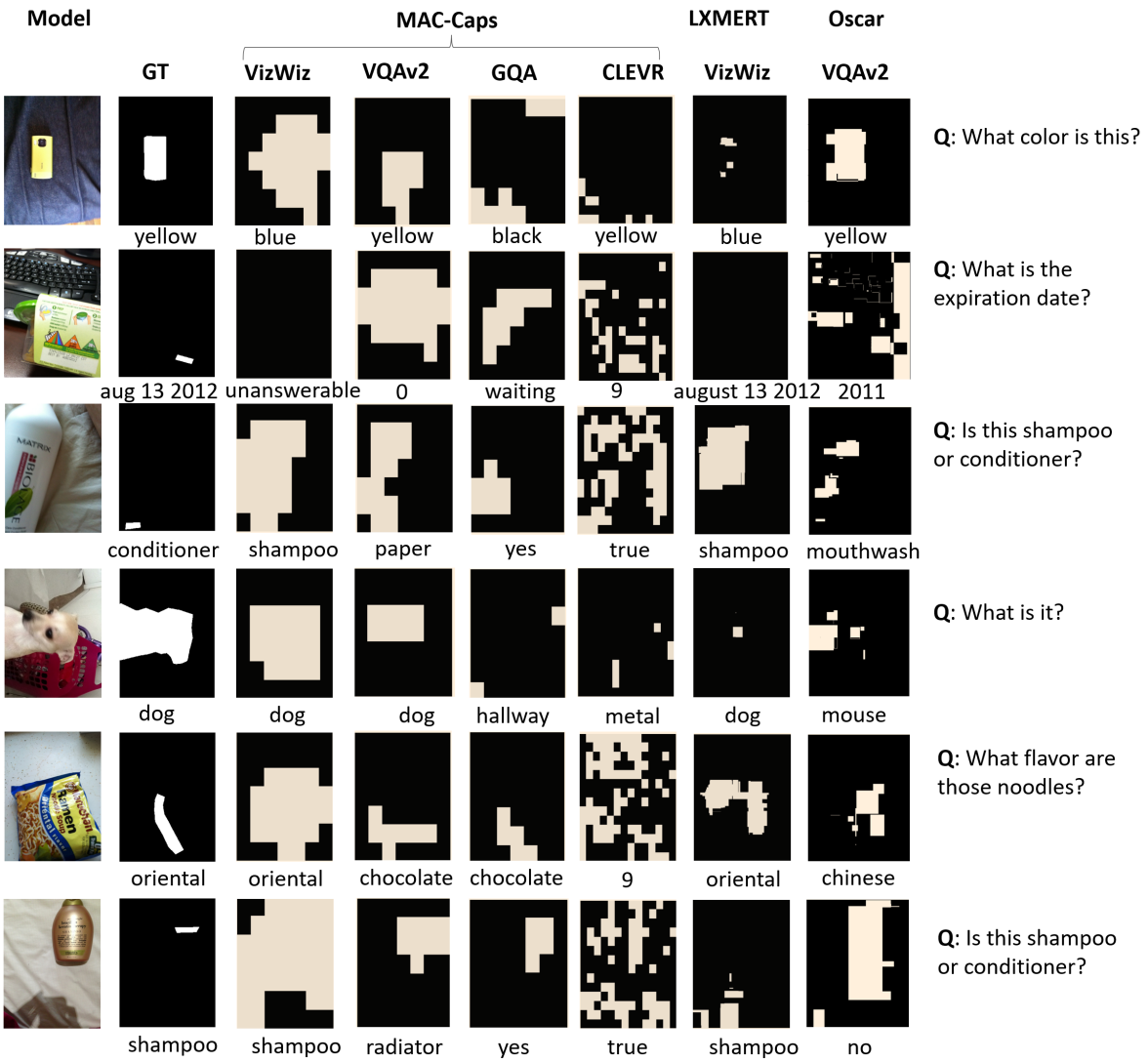

Figure 12. Qualitative results exemplifying answer groundings from 6 models.

Qualitative results. We show examples for the answer groundings predicted by the six benchmarked models. Results are shown in Figure 12. The third column exemplifies the top-performing MAC-Caps model pre-trained on VizWiz. We observe examples for which the MAC-Caps pretrained on VizWiz predicts incorrectly (rows 1-3) and correctly (rows 4-6). Examples highlight that the MAC-Caps model struggles with images containing text (row 2, 3), can correctly predict the answer without grounding the correct region (row 6), and can fail to predict both the correct answer and answer grounding (row 2, 3). We observe for LXMERT that it can similarly predict the answer and answering grounding incorrectly (column 6; row 1, 3) as well as predict the answer correctly while failing to accurately locate the answer grounding (column 7; row 2, 4, 5, 6). We observe for the Oscar model pretrained on VQAv2 that it can detect the foreground object better compared to other models, but fails when the visual evidence is a small region (column 8; row 2, 3).

We found that models (MAC-CAPs and LXMERT) pretrained on VizWiz-VQA failed to answer the question and locate the visual evidence for visual questions requiring color recognition (row 2, column 3,7), while models (MAC-CAPs and OSCAR) pretrained on VQAv2 can answer questions related to color and ground the correct region (row 1, column 4, 8). We found this surprising since the VQAv2 has less visual questions related to color than the VizWiz dataset and also represents a distinct domain [14].

References

[1] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. *arXiv preprint arXiv:2007.14419*, 2020. 8

[2] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. *arXiv preprint arXiv:2003.12511*, 2020. 8

[3] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 8

[4] 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. 8

[5] 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. 8

[6] 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. 8

[7] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 8

[8] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 8

[9] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 8

[10] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable TextVQA models via visual and textual explanations. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 19–29, Mexico City, Mexico, June 2021. Association for Computational Linguistics. 8

[11] 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. 8

[12] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 8

[13] Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8465–8474, 2021. 8

[14] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–31, 2020. 9

[15] 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. 8