Supplementary Materials -
VizWiz Grand Challenge: Answering Visual Questions from Blind People

Danna Gurari[1], Qing Li[2], Abigale J. Stangl[3], Anhong Guo[4], Chi Lin[1],
Kristen Grauman[1], Jiebo Luo[5], and Jeffrey P. Bigham[4]
[1] University of Texas at Austin, [2] University of Science and Technology of China,
[3] University of Colorado Boulder, [4] Carnegie Mellon University [5] University of Rochester

This document supplements the main paper with the following details about:

1. Filtering visual questions (supplements **Section 3.2**).

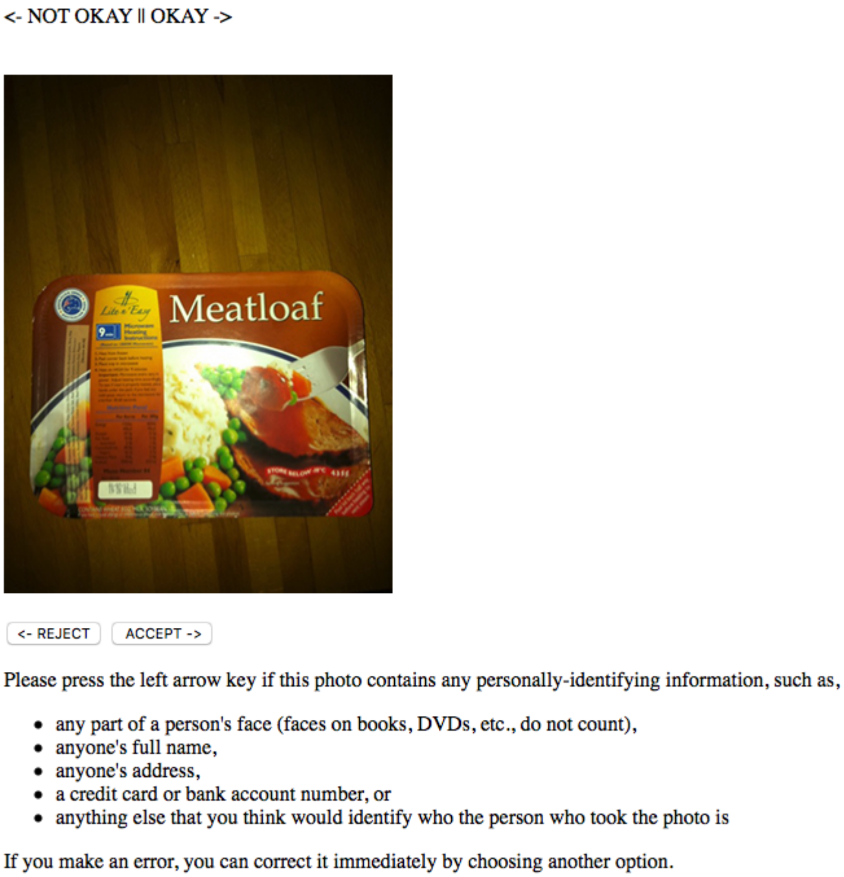
2. Collecting answers to visual questions (supplements **Section 3.3**).

3. Analyzing the VizWiz dataset (supplements **Section 4**).

4. Benchmarking algorithm performance (supplements **Section 5**).

1. Filtering Visual Questions

We first used the crowdsourcing system shown in **Figure 1** to identify images showing personal-identifying information. To err on the safe side in protecting all involved parties, we next iteratively developed a taxonomy of possible vulnerabilities people face when working with a VQA dataset created "in the wild". During an initial brainstorming session, we identified the following three categories: (1) personally-identifying information, also called PII (e.g., any part of a person's face, financial statements, prescriptions), (2) Location (e.g., addressed mail, business locations), and (3) Adult Content (e.g., nudity, cuss words). We then examined the robust-ness of this taxonomy by evaluating the inter-annotator agreement between three domain experts who reviewed 1,000 randomly-selected visual questions and labeled "vulnerable" instances. We found exactly one person marked a visual question for removal for the majority of instances (i.e., 44) that visual questions were tagged for removal (i.e., 64). We found most disagreements occurred on visual questions for which the researchers were not sure, such as in poor quality images or complex scenes. We therefore added two more categories to our taxonomy that reflected our desire to err on the safe side: (4) Questionable Complex Scenes and (5) Questionable Low Quality Images.



Please press the left arrow key if this photo contains any personally-identifying information, such as,

- any part of a person's face (faces on books, DVDs, etc., do not count),
- anyone's full name,
- anyone's address,
- a credit card or bank account number, or
- anything else that you think would identify who the person who took the photo is

If you make an error, you can correct it immediately by choosing another option.

Figure 1: AMT user interface for identifying images showing PII.

2. Answer Collection

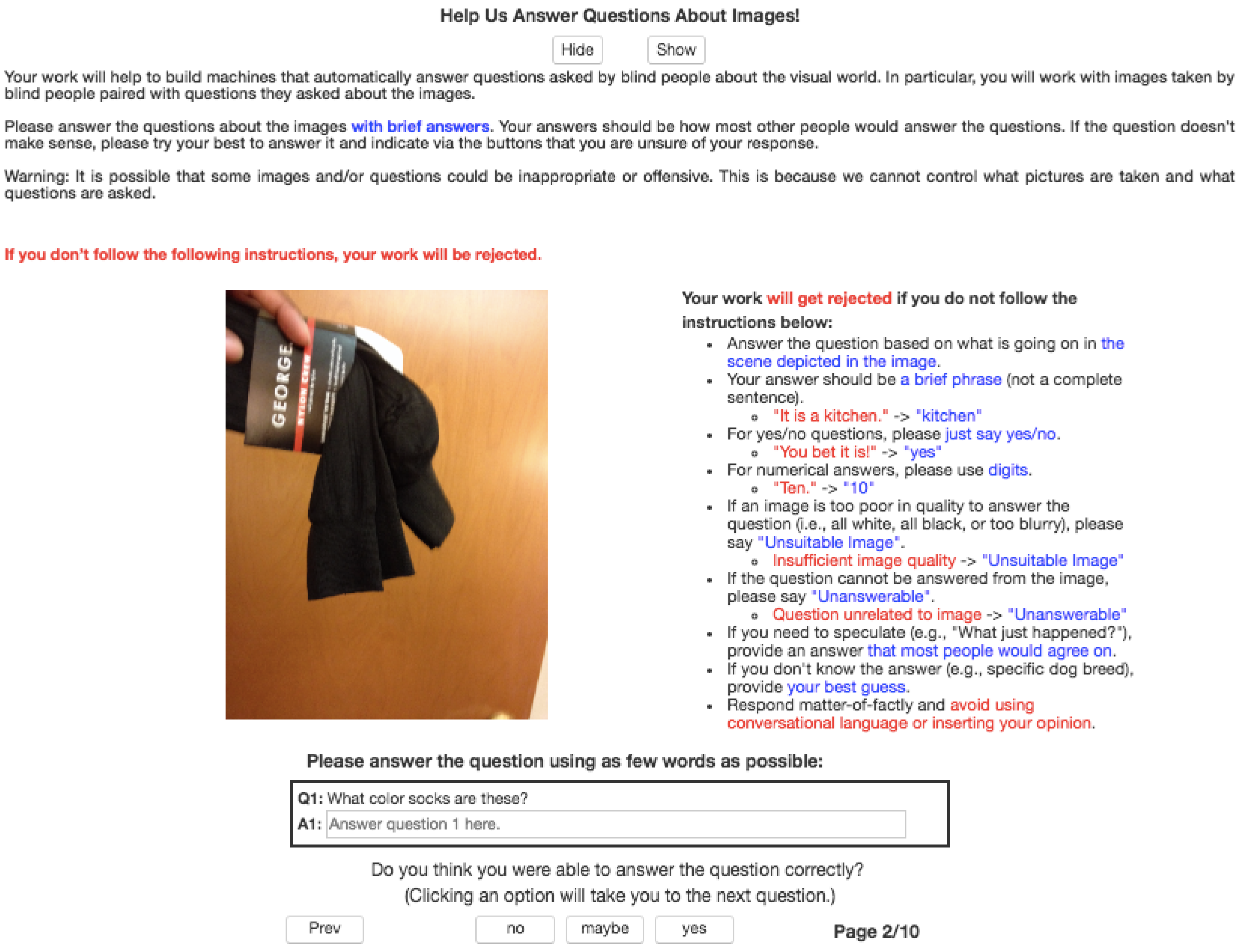
2.1. Answer Post-Processing

Following prior work [2], we converted all letters to lower case, converting numbers to digits, and removing punctuation and articles (i.e., “a”, “an”, “the”). We further post-processed the answers by fixing spelling mistakes and removing filler words (i.e., “it’”, “is”, “its”, “and”, “&”, “with”, “there”, “are”, “of”, “or”). For spell checking, we relied on two automated spell-checkers to reveal which words in the answers neither reflected common nor popular modern words: (1) Enchant[1] provides an API to multiple libraries such as Aspell/Pspell and AppleSpell and (2) an algorithm invented by Google search quality director Peter Norvig[2], that is based on frequent words in popular Wikipedia articles and movie subtitles, and so augments modern words such as iPhone and Gmail. Both the aforementioned tools also employ different mechanisms to return correct word candidates. When the most probable correct word from both tools matched, we replaced the original word with the candidate. For the remaining answers, we solicited the correct spelling of the word from trusted in-house human reviewers. We found many of the detected “misspelled” words were valid captchas and so did not need spell-correction.

2.2. Crowdsourcing System

We show the Amazon Mechanical Turk (AMT) interface that we used to collect answers in **Figure 2**. We limited our users to US citizens to minimize concerns about whether a person is familiar with the language. We also limited our users to those who previously had 95% jobs approved for over 500 jobs to increase the likelihood of collecting high quality results. Finally, we used the “Adult Qualification” in AMT to ensure our selected crowd was comfortable reviewing adult content. This was

[1] https://www.abisource.com/projects/enchant/
[2] http://norvig.com/spell-correct.html



Help Us Answer Questions About Images!

Your work will help to build machines that automatically answer questions asked by blind people about the visual world. In particular, you will work with images taken by blind people paired with questions they asked about the images.

Please answer the questions about the images **with brief answers**. Your answers should be how most other people would answer the questions. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

Warning: It is possible that some images and/or questions could be inappropriate or offensive. This is because we cannot control what pictures are taken and what questions are asked.

If you don't follow the following instructions, your work will be rejected.

Your work will get rejected if you do not follow the instructions below:

- Answer the question based on what is going on in the scene depicted in the image.
- Your answer should be a brief phrase (not a complete sentence).
  - "It is a kitchen." -> "kitchen"
- For yes/no questions, please just say yes/no.
  - "You bet it is!" -> "yes"
- For numerical answers, please use digits.
  - "Ten." -> "10"
- If an image is too poor in quality to answer the question (i.e., all white, all black, or too blurry), please say "Unsuitable Image".
  - Insufficient image quality -> "Unsuitable Image"
- If the question cannot be answered from the image, please say "Unanswerable".
  - Question unrelated to image -> "Unanswerable"
- If you need to speculate (e.g., "What just happened?"), provide an answer that most people would agree on.
- If you don't know the answer (e.g., specific dog breed), provide your best guess.
- Respond matter-of-factly and avoid using conversational language or inserting your opinion.

Please answer the question using as few words as possible:

Q1: What color socks are these?
A1: Answer question 1 here.

Do you think you were able to answer the question correctly?
(Clicking an option will take you to the next question.)

Page 2/10

Figure 2: AMT user interface for collecting answers to visual questions.

important because visual questions are gathered "from the wild" so could contain content that is not appropriate for a general audience (e.g., nudity).

3. VQA Dataset Analysis

3.1. Question Length Distribution

We augment the statistics supplied in the main paper, with the fine-grained distribution showing the number of words in each visual question in **Figure 3**. We cut the plot off at 30 words in the visual question[3]. This distribution highlights the prevalence of outliers with few words or 10s of words in the question.

3.2. Average Image Excluding "Unanswerable" Visual Questions

We show a parallel image supplied in the main paper here, with the only change being that we show the average of all images excluding those coming from visual questions labelled as unanswerable. The resulting image shown in **Figure 4** resembles that shown in the main paper by also being a gray image, and so reflecting a diverse set of images that do not conform to a particular structure.

[3] There is a small tail of visual questions that spread to a maximum of 62 words in the question.

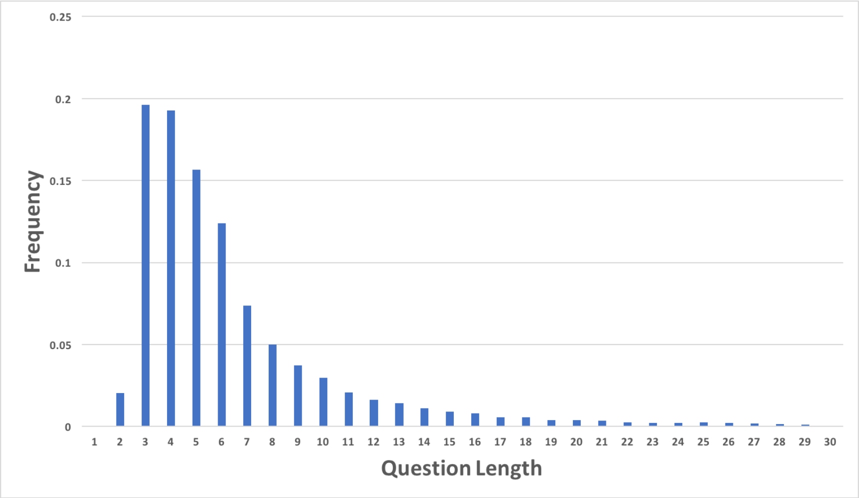
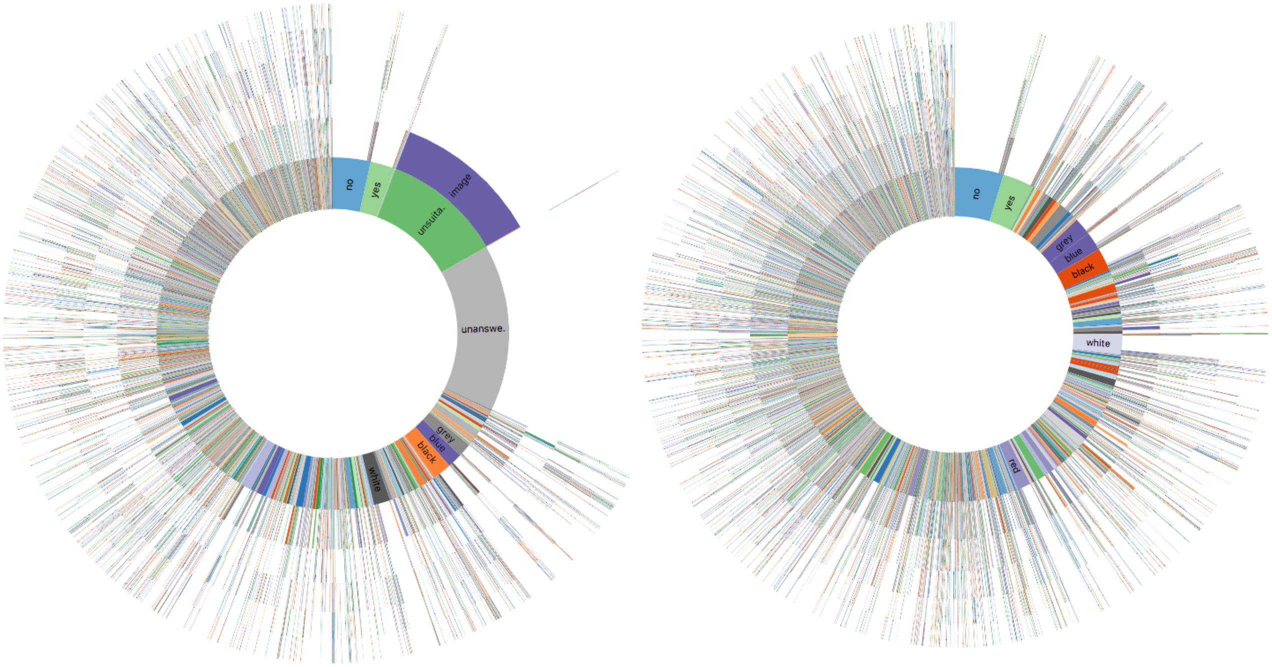


Figure 3: Distribution of number of words per visual question.

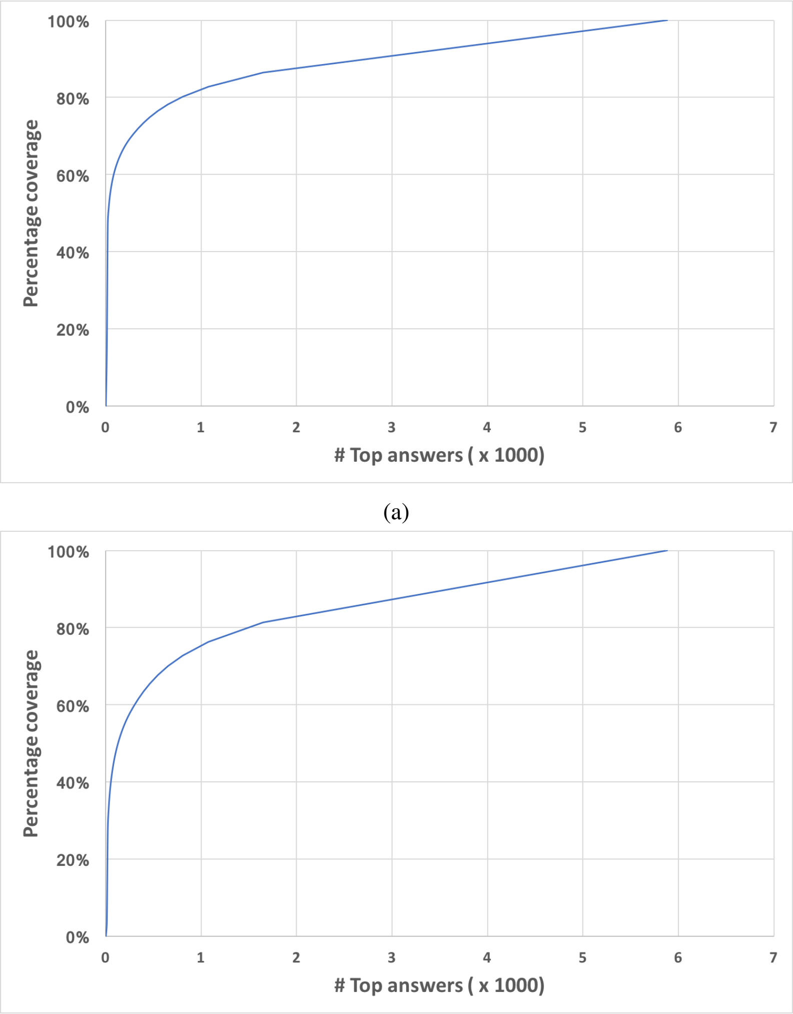
Figure 4: The average image created using all images in VizWiz, excluding those that are in unanswerable visual questions.



(a)

(b)

Figure 5: Distribution of the first six words for (a) all answers in VizWiz and (b) all answers in VizWiz excluding unanswerable visual questions. The innermost ring represents the first word and each subsequent ring represents a subsequent word. The arc size is proportional to the number of answers with that initial word/phrase.



(a)

(b)

Figure 6: Cumulative number of visual questions covered by the most frequent answers in VizWiz for (a) all answers in VizWiz and (b) all answers in VizWiz excluding unanswerable visual questions.

3.3. Answer Analysis

We show in **Figure 5** sunburst diagrams which visualize the frequency that answers begin with different words/phrases. The most common answers, following "Unsuitable Image" and "Unanswerable", are yes, no, and colors. We observe there is a large diversity of uncommon answers as well as answer lengths spanning up to 6 words long.

We also show in **Figure 6** plots of the cumulative coverage of all answers versus the most frequent answers. The straight line with a slope of roughly 1 further illustrates the prevalence of a long tail of unique answers.

4. VizWiz Algorithm Benchmarking

4.1. VQA

In the main paper, we report results for the algorithm proposed in [1]. We use our implementation since the authors' code is not yet publicly-available. In particular, the authors released their code and pre-trained faster RCNN model for the image feature extraction, but not the code and the pre-trained model for the VQA part.

In the main paper, we also report results for fine-tuned models. We fine-tune each pre-trained model on VizWiz using the most frequent 3,000 answers in the training set of VizWiz. For the initialization of the last layer, if the answer is in the candidate answer set of VQA V2.0 dataset [3], we initialize the corresponding parameters from the pre-trained model, and if not, we randomly initialize the parameters. We use Adam solver [6] with a batch size of 128 and an initial learning rate of 0.01 that is dropped to 0.001 after the first 10 epochs. The training is stopped after another 10 epochs. We employ both dropout [8] and batch normalization [4] during training.

In the main paper, we also report results for models trained from scratch. Each model is trained using the 3,000 most frequent answers in the train split of VizWiz. We initialize all parameters in the model to random values.

Finally, we report fine-grained details to expand on our findings reported in the main paper about how well algorithms trained on VizWiz predict answers for the VQA 2.0 test dataset [3]. We report results for the six models that are fine-tuned and trained from scratch for the three models [1, 3, 5] with respect to all visual questions as well as with respect to the four answer types in **Table 1**. These results highlight that VizWiz provides a domain shift to a different, difficult VQA environment compared to existing datasets.

	All	Yes/No	Number	Other
FT [3]	0.300	0.612	0.094	0.079
FT [5]	0.318	0.601	0.163	0.110
FT [1]	0.304	0.595	0.082	0.105
VizWiz [3]	0.218	0.461	0.074	0.042
VizWiz [5]	0.228	0.465	0.131	0.049
VizWiz [1]	0.219	0.453	0.083	0.048

Table 1: Shown is the cross-dataset performance of six models trained on VizWiz and tested on the VQA 2.0 test dataset [3].

4.2. Answerability

Below is a brief description of the implementations of the models we use in the main paper:

- `Q`: a one-layer LSTM is used to encode the question and is input to a softmax layer.
- `C`: a one-layer LSTM is used to encode the caption and is input to a softmax layer.
- `I`: ResNet-152 is used to extract the image features from the pool5 layer and is input to a softmax layer.
- `Q+C`: the question and caption are encoded by two separate LSTMs and then the features of the question and caption are concatenated and input to a softmax layer.
- `Q+I` the features of question and image are concatenated and input to a softmax layer.

For the fine-tuned model, we initialize the parameters using the pre-trained model. We train from scratch by randomly initializing the parameters. For both approaches, we train for 10 epochs on the VizWiz dataset.

We augment here our findings of the average precision in the main paper with the average F1 score in **Table 2**. As observed, the top-performing method remains `Q+I` whether using the AP score or F1 score.

We also show the top 10 most confident answerable and answerable predictions for the top-performing `Q+I` implementation in **Figure 7**. Our findings highlight how predictive cues may relate to the quality of images and specific questions (e.g., “What color...?”).

Model	Average Precision	Average F1 score
Q+C [7]	0.306	0.383
FT [7]	0.561	0.542
VizWiz [7]	0.605	0.549
VQA [3]	0.560	0.569
Q	0.490	0.233
C	0.464	0.270
I	0.640	0.518
Q+I	0.717	0.648

Table 2: Shown are the average precision scores and average F1 scores for eight models used to predict whether a visual question is answerable.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017. 5, 6

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 2

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*, 2016. 6, 7

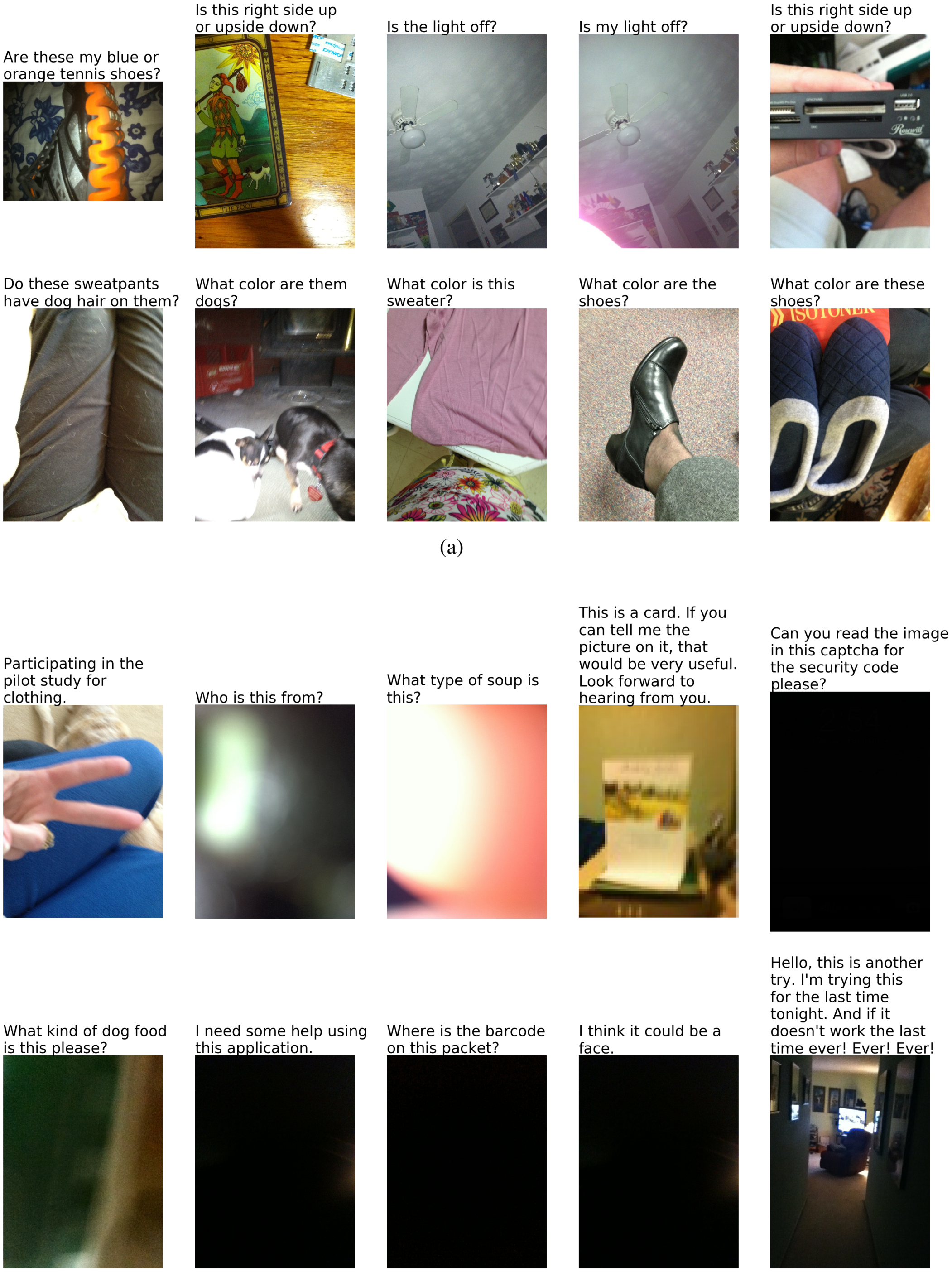
[4] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 6

[5] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 6

[6] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[7] A. Mahendru, V. Prabhu, A. Mohapatra, D. Batra, and S. Lee. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601*, 2017. 7

[8] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. 6


Are these my blue or orange tennis shoes?
Is this right side up or upside down?
Is the light off?
Is my light off?
Is this right side up or upside down?
Do these sweatpants have dog hair on them?
What color are them dogs?
What color is this sweater?
What color are the shoes?
What color are these shoes?
ISOTONER
Participating in the pilot study for clothing.
Who is this from?
What type of soup is this?
This is a card. If you can tell me the picture on it, that would be very useful. Look forward to hearing from you.
Can you read the image in this captcha for the security code please?
What kind of dog food is this please?
I need some help using this application.
Where is the barcode on this packet?
I think it could be a face.
Hello, this is another try. I'm trying this for the last time tonight. And if it doesn't work the last time ever! Ever! Ever!

(a)

(b)

Figure 7: Top 10 most confident predictions by the top-performing Q+I model for visual questions in the VizWiz test dataset that are (a) answerable and (b) unanswerable.