Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models – Supplementary Material

adversarialvqa.github.io

Linjie Li[1], Jie Lei[2], Zhe Gan[1], Jingjing Liu[3]
[1]Microsoft [2]UNC Chapel Hill [3]Tsinghua University
{lindsey.li, zhe.gan}@microsoft.com
jielei@cs.unc.edu, JJLiu@air.tsinghua.edu.cn

A. More Discussions on AVQA

Future Practices. We recommend future models to report performance on both VQA v2 [4, 7] and AVQA. AVQA is designed to test VQA model robustness under human adversarial attacks. It is complementary to VQA v2 (naturally-collected questions), rather than a replacement. In addition, we believe it is beneficial to evaluate on other robust VQA benchmarks as well. While AVQA encompasses broader robustness types and image domains with higher data quality, existing robustness benchmarks [2, 1, 6, 12] can in addition provide useful analysis tailored to individual robustness types. An ideal VQA system should perform well on all VQA benchmarks. Further, we encourage future work to apply human-in-the-loop adversarial attack to their proposed models to identify potential vulnerabilities. For AVQA, we expect to provide a dynamically evolving VQA benchmark as models grow more robust, to alleviate the drawbacks of static benchmarks (*e.g.*, performance saturation and overfitting).

Constraints/Rules For Data Collection. As our goal is to examine VQA models' robustness when encountering test examples in the wild, we do not constrain the questions to specific types, to avoid unconscious bias from dataset creators. As a result, we have found that models make more mistakes on Count/OCR/Relation/Commonsense questions (Table 7 in the main text).

To obtain high-quality adversarial questions, we enforce a set of rules to ensure the questions are objective, relevant to the image, and have exact answers (see detailed instructions in Figure 13). We also manually filter out questions with repetitive patterns for each annotator during collection. Our answer annotation process validates the collected questions to some extent, which are answerable by human but not always answerable by model.

Bias in Model Choices. Our current choice of models was guided by the assumption that newer models are more likely to be transformer-based with currently-proven most effective features. We plan to include a broader choice of models in future collection, as the benchmark evolves.

B. Data Statistics

Type of Questions. Following [4], given the structure of questions generated in English, we cluster questions into different types based on the words that start the question. Figure 1 shows the distribution of questions based on the first four words of the questions in AVQA. Interestingly, the variety of question types are quite similar to those in [4], including “What is”, “How many” and “Is there”. Quantitatively, we also categorize the questions into “Y/N”, “Num”, “OOV” and “Other”. The percentage of questions for different categories is shown in Table 1. “OOV” questions refer to questions that cannot be answered by VQA v2 [7] answer vocabularies. We also include two upper bounds, one based on VQA v2 answer vocabularies, and the other on open vocabularies. Moreover, we estimate human performance on AVQA by sampling 1 human answer as prediction and use the rest 9 answers as references. We repeat the process 10 times and average the score. Comparing model performance reported in the main text, there is still a huge gap, with about 50 points lower than the upper bounds or the estimated human performance.

Question Lengths. Figure 2 shows the distribution of question lengths. We see that most questions range from four to ten words.

Dataset Properties Across Rounds. Figure 3 shows a histogram of the number of tries per verified example across the three different rounds. We observe a consistent trend for all three rounds, over 80% of examples are successfully collected within 2 tries. Figure 4 shows the time taken per verified example. As the round progresses, we observe that more and more examples are collected within 100 seconds (less than 2 minutes). Figure 5 shows the propor-

Round	Question Types				Upper Bound		Human Performance
	Y/N	Num	OOV⋆	Other	dev/test⋆	dev/test†	dev/test
R1	13.53%	23.36%	10.03%	50.08%	81.43/79.75	92.03/92.05	74.92/75.14
R2	8.62%	29.91%	14.37%	47.01%	76.26/77.11	93.60/93.43	78.29/78.83
R3	11.24%	35.55%	12.17%	41.04%	79.64/80.91	94.48/94.41	81.61/81.15
AVQA	11.40%	28.90%	11.95%	47.75%	79.27/79.28	93.28/93.21	78.05/78.15

Table 1: Question type distribution on verified examples and upper bound on dev/test set across three rounds. ⋆ is based on VQA v2 [7] answer vocabularies. † is based on open vocabularies.

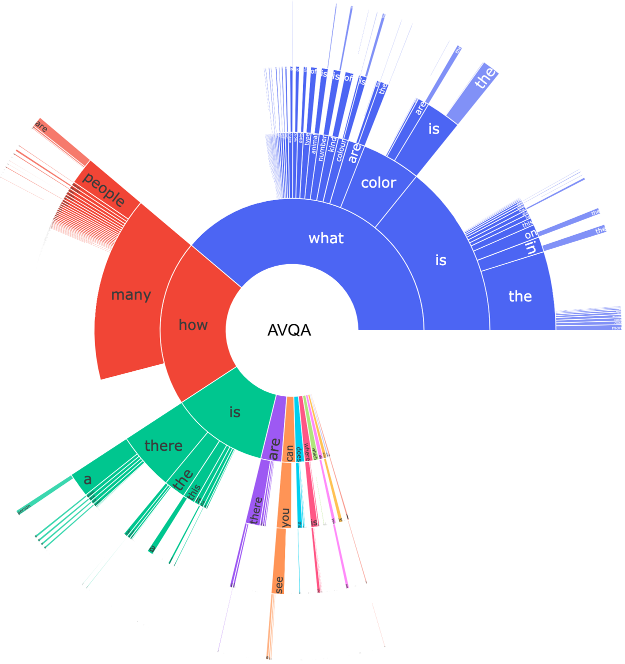


Figure 1: Distribution of questions by their first four words. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

tion of different types of collected examples across three rounds. Comparing to R1 and R2, R3 contains more "not sure" judgements to model answers during question collection (type **B**), which indicates that the task is getting harder. There are a small amount of examples in all three rounds that there is no agreement among the answers collected (type **D**). Examples from **B** an **D** are excluded due to low quality. The rest are split into train/dev/test set (refer to Figure 5 captions for more information).

Answer Confidence and Inter-human Agreement. During answer collection (see interface in Figure 12), the annotators are required to provide both a correct answer to the question given the image content and a self-judgment on how confident they feel about the answer. Specifically, we ask "Do you think you were able to answer the question correctly?", and the annotator need to choose from "yes" (confident with score 1), "maybe" or "no" (not confident with score 0). Figure 6 shows the distribution of responses (black lines). A majority of the answers were labeled as confident. More than 9 annotators are confident about their answers on over 60% questions on average.

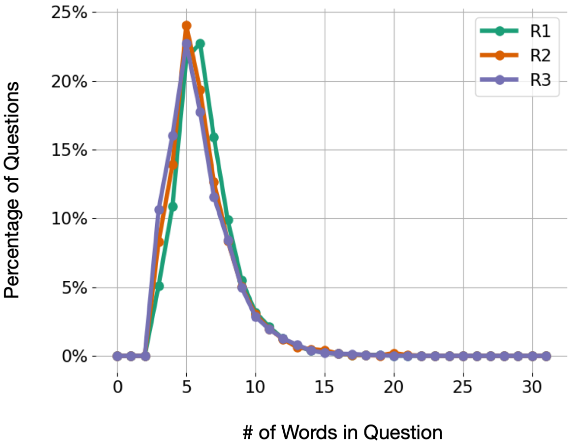


Figure 2: Percentage of questions with different word length across three rounds. Most questions range from four to ten words.

In addition, we investigate how the self-judgment confidence corresponds to the answer agreement between annotators across three rounds of data collection. Color bars in Figure 6 show the percentage of questions in which (i) 7 or more, (ii) 3-7, or (iii) less than 3 workers agree on the answers given their average confidence score. Across all rounds, the agreement between subjects increases with confidence. We do observe that workers are more confident about their answers in R2 and R3, comparing to R1.

Answer Distribution. Figure 7 shows the distribution of answers for several question types. We can see that a number of question types, such as "Is . . . ", "Can. . . ", and "Does. . . " are typically answered using "yes" and "no" as answers. Other questions such as "What is/are. . . " and "What kind/type. . . " have a rich diversity of responses. Other question types such as "What color. . . " or "Which. . . " have more specialized responses, such as colors, or "left" and "right". These observations are similar to those in VQA v2.

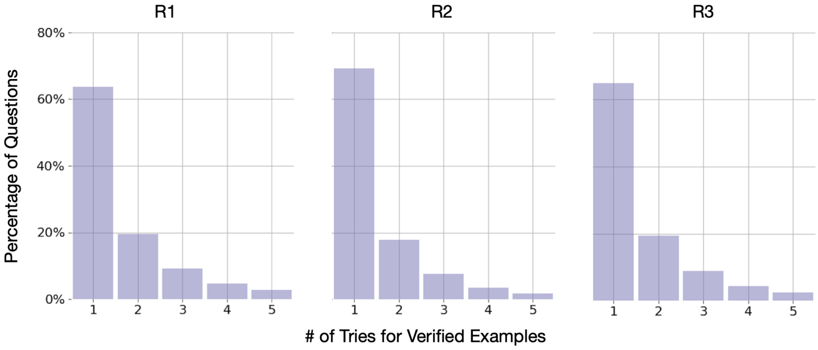


Figure 3: Histogram of the number of tries for each good verified example across three rounds.

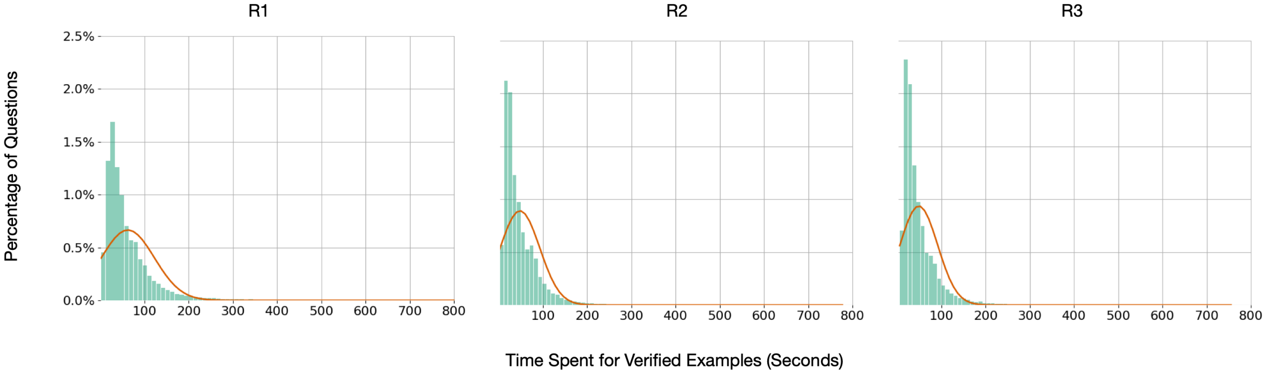


Figure 4: Histogram of the time spent per good verified example across three rounds.

C. More Visualizations

We include more visualization examples of collected data across three rounds in Figure 8. We show adversarial questions from 4 categories: Count, OCR, Reasoning and Visual Concept Recognition. Note that questions may belong to multiple categories. For example, counting question from R3 ("How many natural satellites are in the sky?") requires commonsense about "natural satellites". OCR question from R1 ("What company is on the back of the referee?") not only requires commonsense about "referee" but relational reasoning about "on the back of". Reasoning questions include positional/relational reasoning (*e.g.*, "What is the woman closest to the camera holding in her hand?"), commonsense reasoning (*e.g.*, "Is the egg yolk cooked?") and comparative reasoning ("Who is taller?"). There are also questions that require recognition of both low-level visual concepts (*e.g.*, color/shape) and high-level visual concepts (*e.g.*, action, relation).

We also visualize more examples generated via textual adversarial attack methods (Sears [11], Textfooler [8] and Sememe+PSO [14]) in Figure 9. The first two columns show invalid examples, and the last column includes valid examples, based on our manual examination. Recall that our goal is to collect high-quality adversarial questions that can be used to *accurately, thoroughly evaluate and examine* the weakness of VQA models. Automatically generated adversarial questions are often incorrect (requiring additional human efforts to validate their correctness), and limited to linguistic variations to existing questions, thereby they are unlikely to provide a comprehensive analysis.

D. More Results

Recall that questions in R3 are collected on images from various domains, including web images from Conceptual Captions [13] (CC, used in R1 and R2), user-generated images from Fakeddit [10] and movie video frames from VCR [15]. Hence, we can study how model performance can be transferable across different domains. We create a new split of R3 (R3*) according to the image source, with CC images for training and Fakeddit/VCR images for evaluation. Table 2 summarizes UNITER-B performance under different training settings. Despite the domain differences in images, the performance on Fakeddit and VCR split improves as we include more training data from CC images. Comparing the new split R3* with the original split R3, training on more in-domain examples on CC images does help to improve model performance on R1 and R2.

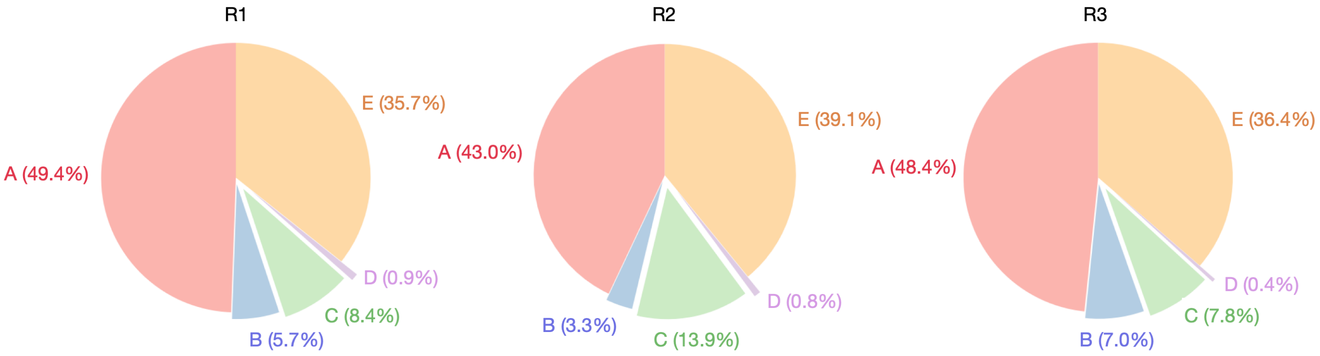


Figure 5: Proportion across three rounds. **A**=Examples that model got right ("Definitely Correct") during question collection, **B**=Examples that model neither got right nor wrong ("Not Sure") during question collection. **C**, **D** and **E** are examples that model got wrong ("Definitely Wrong") during question collection and sent to 9 annotators for verification during answer collection. Specifically, **C**=Examples that more than 3 verifiers overruled the question author's decision of "Definitely Wrong" and agree with the model's answer. **D**=Examples for which there is no agreement among verifiers, **E**=Examples where at least two verifiers agree with each other during answer collection. We split **E** by images into training, dev, and test sets. Examples on training images in **A** and **C** are added to the training set, the rest are discarded. **B** and **D** are excluded due to low quality.

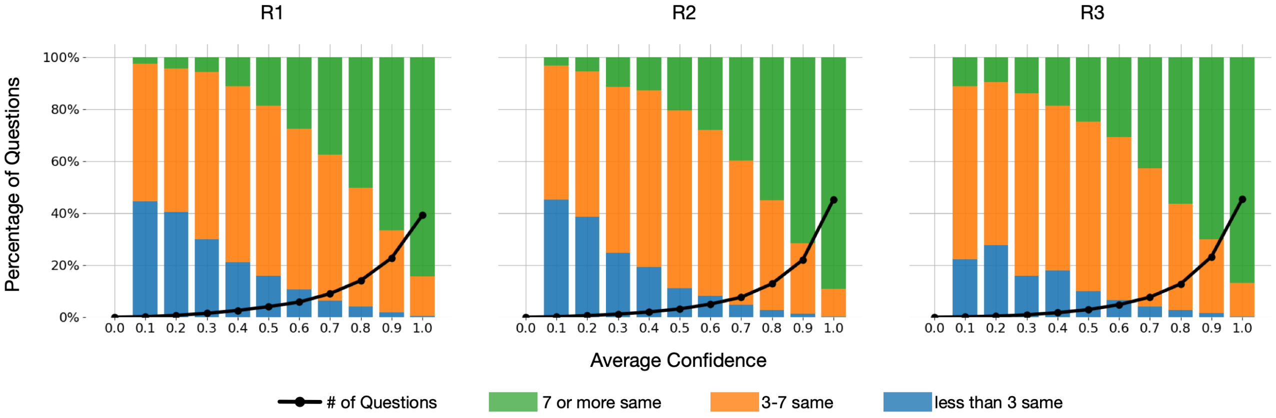


Figure 6: Number of questions per average confidence score across three rounds (black lines, 0 = not confident, 1 = confident). Percentage of questions where 7 or more answers are same, 3-7 are same, less than 3 are same across three rounds (color bars).

Training Data	R1	R2	R3	Fakeddit [10]	VCR [15]
VQA v2+VGQA	20.60	17.86	20.71	19.59	23.34
+R1	26.03	17.30	20.56	20.27	23.84
+R2	26.60	23.21	19.26	17.85	22.05
+R3*	**27.02**	**23.78**	-	**22.56**	**27.43**
ALL	26.85	23.38	**24.48**	-	-

Table 2: Domain transfer evaluation on UNITER-B. $\star$ indicates that we only use examples collected on CC [13] images for training. ALL refers to VQA v2+VGQA+R1+R2+R3.

We also observe that model performance on VCR is significantly higher than those on the original R3 dev and Fakeddit splits across all training settings. Images from VCR are often human-centric, which may be "easier" than complex or abstract scenes depicted in CC/Fakeddit images.

In addition, we include detailed results from BUTD [3], ClipBERT [9], VILLA-B and VILLA-L [5] in Table 3. These results are consistent with observations we summarized in Section 4 of the main text.

E. Data Collection Interface

Examples of the user interface are shown in Figures 10, 11 and 12. We also include full instructions and examples shown to the annotators in Figures 13 and 14.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020. 1

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 1

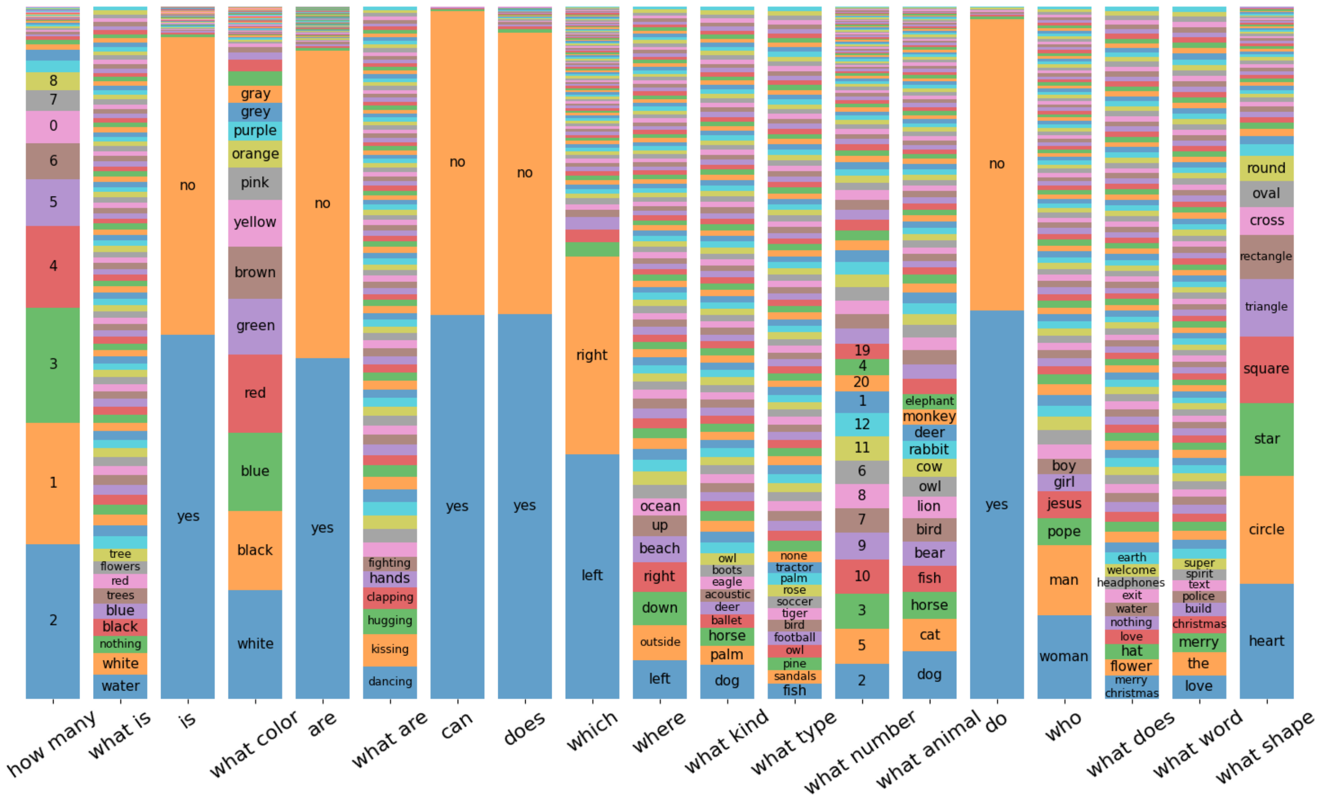


Figure 7: Distribution of answers per question type. Only top-100 answers to each question type are plotted. The height of each color bar is proportional to the percentage of an answer to the corresponding question type.

Model	Training Data	R1	R2	R3	AVQA	VQA v2	Δ(v2, AVQA)
		dev/test	dev/test	dev/test	dev/test	test-dev	test-dev, test
BUTD	VQA v2 +VGQA	20.80/19.28	18.77/18.85	20.63/21.10	20.12/19.71	67.60	47.89
	+R1	20.27/20.27	19.53/20.14	21.55/21.86	20.44/20.72	67.37	46.65
	+R1+R2	24.41/21.82	22.28/21.80	21.31/21.60	22.78/21.75	67.44	45.69
	ALL	24.96/22.11	22.62/22.78	23.92/23.61	23.91/22.78	67.52	**44.74**
ClipBERT	VQA v2 +VGQA	21.39/20.45	19.29/20.06	21.01/23.16	20.45/21.16	69.08	47.92
	+R1	23.83/22.43	20.08/20.13	22.49/22.65	22.25/21.78	69.07	47.29
	+R1+R2	24.03/23.08	23.12/23.86	24.67/23.37	23.95/23.86	69.19	45.33
	ALL	24.62/23.68	22.96/24.66	**25.05/24.87**	24.24/24.35	69.17	44.82
VILLA-B	VQA v2 +VGQA	21.22/19.45	18.53/18.92	20.57/20.73	20.18/19.68	73.37	53.69
	+R1	25.92/24.07	20.00/20.05	21.61/21.23	22.74/21.93	73.21	51.28
	+R1+R2	27.53/25.13	23.23/23.91	21.96/21.87	24.46/23.74	73.11	49.37
	ALL	**30.78/28.43**	**25.66/25.11**	24.00/24.18	**27.08/26.08**	74.28	48.20
VILLA-L	VQA v2 +VGQA	24.99/22.88	18.58/18.23	20.07/19.64	21.47/20.42	**74.58**	54.16
	+R1	28.29/26.12	19.44/19.02	20.25/20.25	23.04/22.08	74.12	52.04
	+R1+R2	30.02/27.81	24.05/23.59	19.38/20.50	24.85/24.24	74.06	49.82
	ALL	29.92/28.01	24.59/24.26	23.66/23.09	26.32/25.32	74.24	48.92

Table 3: Detailed results from BUTD [3], ClipBERT [9], VILLA-B and VILLA-L [5] under different settings. AVQA = R1+R2+R3, ALL = VQA v2+VGQA+AVQA.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 4, 5

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1

[5] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng,

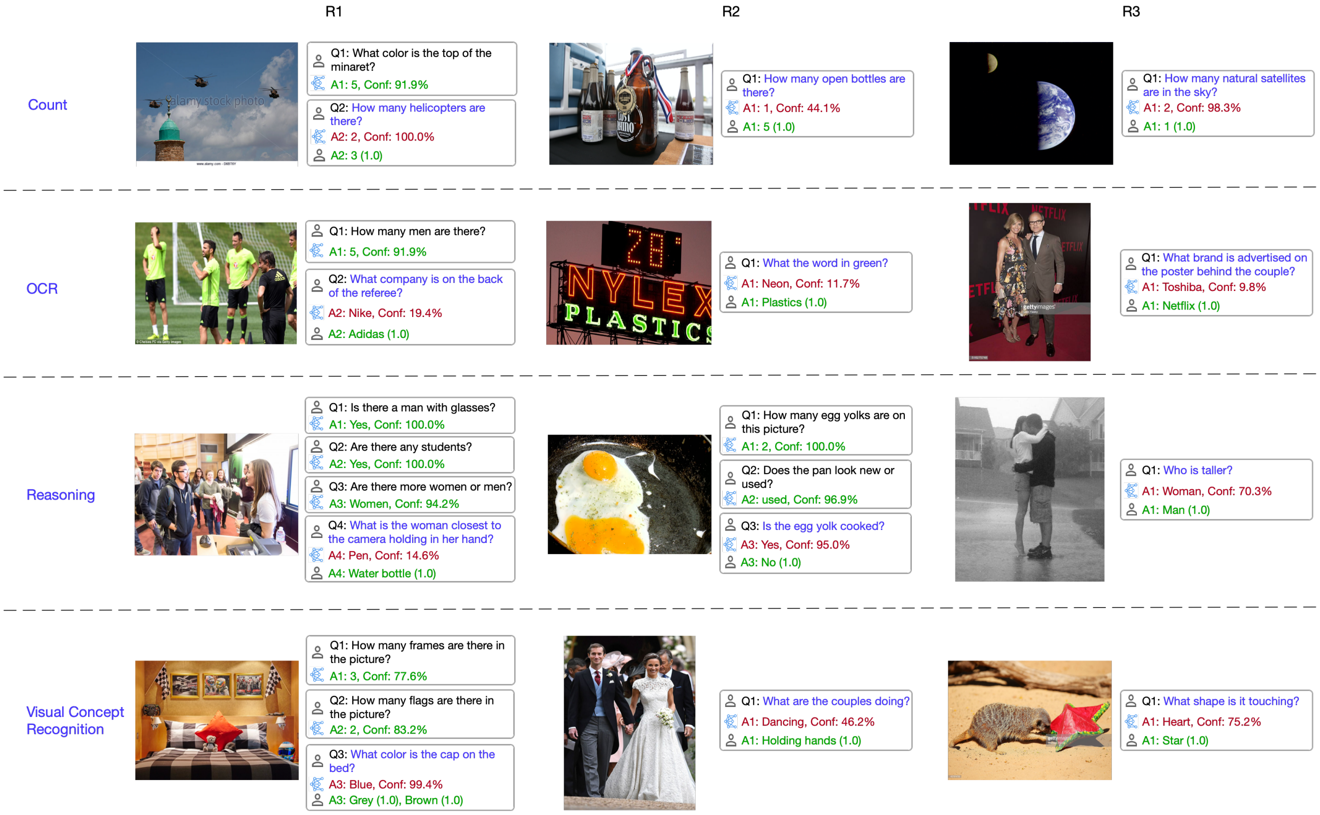

R1
R2
R3
Count
Q1: What color is the top of the minaret?
A1: 5, Conf: 91.9%
Q2: How many helicopters are there?
A2: 2, Conf: 100.0%
A2: 3 (1.0)
Q1: How many open bottles are there?
A1: 1, Conf: 44.1%
A1: 5 (1.0)
Q1: How many natural satellites are in the sky?
A1: 2, Conf: 98.3%
A1: 1 (1.0)
OCR
Q1: How many men are there?
A1: 5, Conf: 91.9%
Q2: What company is on the back of the referee?
A2: Nike, Conf: 19.4%
A2: Adidas (1.0)
Q1: What the word in green?
A1: Neon, Conf: 11.7%
A1: Plastics (1.0)
Q1: What brand is advertised on the poster behind the couple?
A1: Toshiba, Conf: 9.8%
A1: Netflix (1.0)
Reasoning
Q1: Is there a man with glasses?
A1: Yes, Conf: 100.0%
Q2: Are there any students?
A2: Yes, Conf: 100.0%
Q3: Are there more women or men?
A3: Women, Conf: 94.2%
Q4: What is the woman closest to the camera holding in her hand?
A4: Pen, Conf: 14.6%
A4: Water bottle (1.0)
Q1: How many egg yolks are on this picture?
A1: 2, Conf: 100.0%
Q2: Does the pan look new or used?
A2: used, Conf: 96.9%
Q3: Is the egg yolk cooked?
A3: Yes, Conf: 95.0%
A3: No (1.0)
Q1: Who is taller?
A1: Woman, Conf: 70.3%
A1: Man (1.0)
Visual Concept Recognition
Q1: How many frames are there in the picture?
A1: 3, Conf: 77.6%
Q2: How many flags are there in the picture?
A2: 2, Conf: 83.2%
Q3: What color is the cap on the bed?
A3: Blue, Conf: 99.4%
A3: Grey (1.0), Brown (1.0)
Q1: What are the couples doing?
A1: Dancing, Conf: 46.2%
A1: Holding hands (1.0)
Q1: What shape is it touching?
A1: Heart, Conf: 75.2%
A1: Star (1.0)

Figure 8: More visualization of examples collected per round in AVQA. We show examples that contains adversarial questions from 4 categories: Count, OCR, Reasoning and Visual Concept Recognition across three rounds. Each ground-truth answer (VQA score) is collected from 10 workers. Green (red) indicates a correct (wrong) answer. Blue highlights the verified adversarial questions.

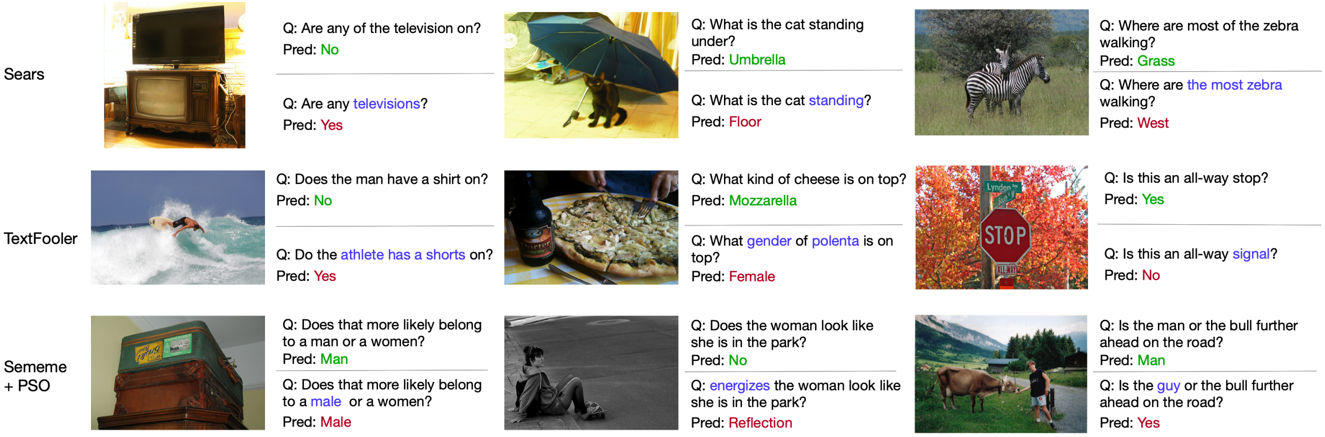

Sears
Q: Are any of the television on?
Pred: No
Q: Are any televisions?
Pred: Yes
Q: What is the cat standing under?
Pred: Umbrella
Q: What is the cat standing?
Pred: Floor
Q: Where are most of the zebra walking?
Pred: Grass
Q: Where are the most zebra walking?
Pred: West
TextFooler
Q: Does the man have a shirt on?
Pred: No
Q: Do the athlete has a shorts on?
Pred: Yes
Q: What kind of cheese is on top?
Pred: Mozzarella
Q: What gender of polenta is on top?
Pred: Female
Q: Is this an all-way stop?
Pred: Yes
Q: Is this an all-way signal?
Pred: No
Sememe + PSO
Q: Does that more likely belong to a man or a women?
Pred: Man
Q: Does that more likely belong to a male or a women?
Pred: Male
Q: Does the woman look like she is in the park?
Pred: No
Q: energizes the woman look like she is in the park?
Pred: Reflection
Q: Is the man or the bull further ahead on the road?
Pred: Man
Q: Is the guy or the bull further ahead on the road?
Pred: Yes

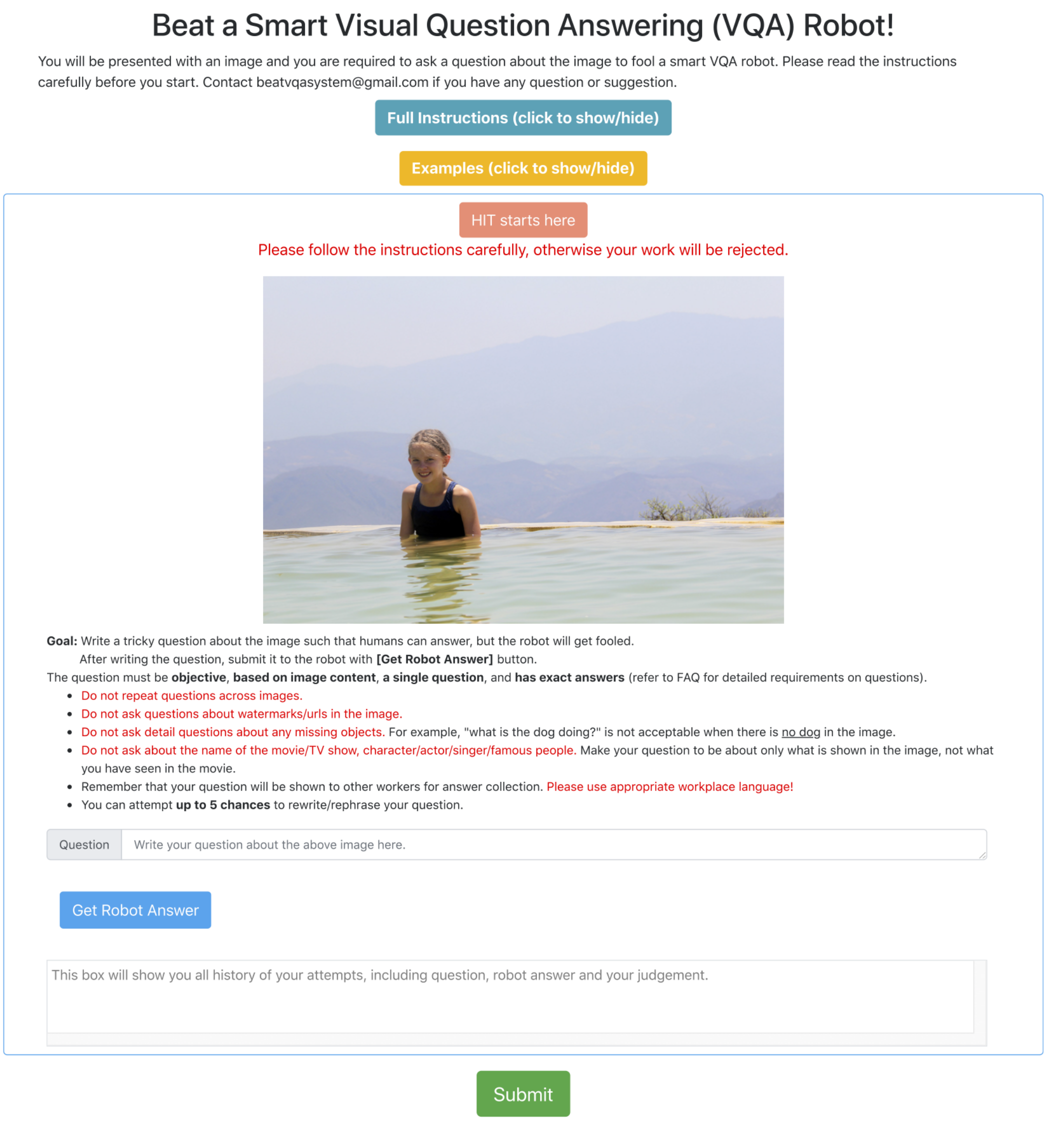
Figure 9: More adversarial examples from textual adversarial attack methods: Sears [11], Textfooler [8] and Sememe+PSO [14]. Green (red) indicates a correct (wrong) answer. Blue highlights the changes made in adversarial questions.

and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *NeurIPs*, 2020. 4, 5

[6] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. *ECCV*, 2020. 1

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1, 2

[8] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, 2020.



Beat a Smart Visual Question Answering (VQA) Robot!

You will be presented with an image and you are required to ask a question about the image to fool a smart VQA robot. Please read the instructions carefully before you start. Contact beatvqasystem@gmail.com if you have any question or suggestion.

Full Instructions (click to show/hide)
Examples (click to show/hide)
HIT starts here

Please follow the instructions carefully, otherwise your work will be rejected.

Goal: Write a tricky question about the image such that humans can answer, but the robot will get fooled.
After writing the question, submit it to the robot with **[Get Robot Answer]** button.
The question must be **objective**, **based on image content**, **a single question**, and **has exact answers** (refer to FAQ for detailed requirements on questions).

- Do not repeat questions across images.
- Do not ask questions about watermarks/urls in the image.
- Do not ask detail questions about any missing objects. For example, "what is the dog doing?" is not acceptable when there is no dog in the image.
- Do not ask about the name of the movie/TV show, character/actor/singer/famous people. Make your question to be about only what is shown in the image, not what you have seen in the movie.
- Remember that your question will be shown to other workers for answer collection. Please use appropriate workplace language!
- You can attempt **up to 5 chances** to rewrite/rephrase your question.

Question: Write your question about the above image here.

This box will show you all history of your attempts, including question, robot answer and your judgement.

Figure 10: UI for question collection. Given an image, the annotator is required to write a tricky answer to fool our "smart VQA robot" (well-trained VQA models). After clicking the "Get Robot Ansner", the annotated question will be sent to our online model for evaluation, and a feedback will be returned immediately. See Figure 11 for an example of model feedback.

3, 6

[9] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 4, 5

[10] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019. 3, 4

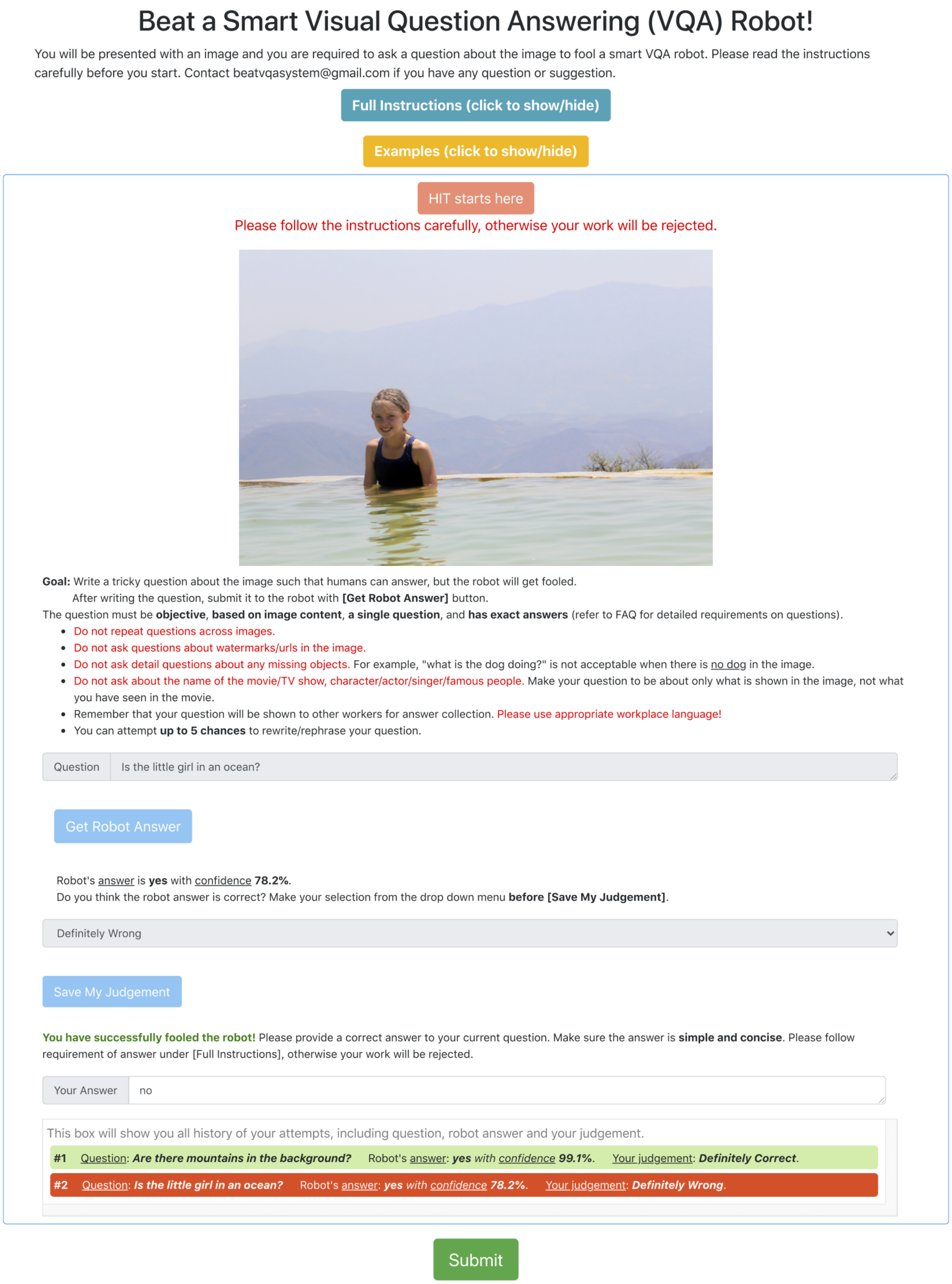
[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, 2018. 3, 6

[12] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Interrogating vqa models with sub-questions. *CVPR*, 2020. 1

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3, 4

[14] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. In *ACL*, 2020. 3, 6

[15] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reason-



Beat a Smart Visual Question Answering (VQA) Robot!

You will be presented with an image and you are required to ask a question about the image to fool a smart VQA robot. Please read the instructions carefully before you start. Contact beatvqasystem@gmail.com if you have any question or suggestion.

Full Instructions (click to show/hide)
Examples (click to show/hide)
HIT starts here

Please follow the instructions carefully, otherwise your work will be rejected.

Goal: Write a tricky question about the image such that humans can answer, but the robot will get fooled.
After writing the question, submit it to the robot with **[Get Robot Answer]** button.
The question must be **objective**, **based on image content**, **a single question**, and **has exact answers** (refer to FAQ for detailed requirements on questions).

- Do not repeat questions across images.
- Do not ask questions about watermarks/urls in the image.
- Do not ask detail questions about any missing objects. For example, "what is the dog doing?" is not acceptable when there is no dog in the image.
- Do not ask about the name of the movie/TV show, character/actor/singer/famous people. Make your question to be about only what is shown in the image, not what you have seen in the movie.
- Remember that your question will be shown to other workers for answer collection. Please use appropriate workplace language!
- You can attempt **up to 5 chances** to rewrite/rephrase your question.

Question: Is the little girl in an ocean?

Robot's answer is **yes** with confidence **78.2%**.
Do you think the robot answer is correct? Make your selection from the drop down menu **before [Save My Judgement]**.

Definitely Wrong

You have successfully fooled the robot! Please provide a correct answer to your current question. Make sure the answer is **simple and concise**. Please follow requirement of answer under [Full Instructions], otherwise your work will be rejected.

Your Answer: no

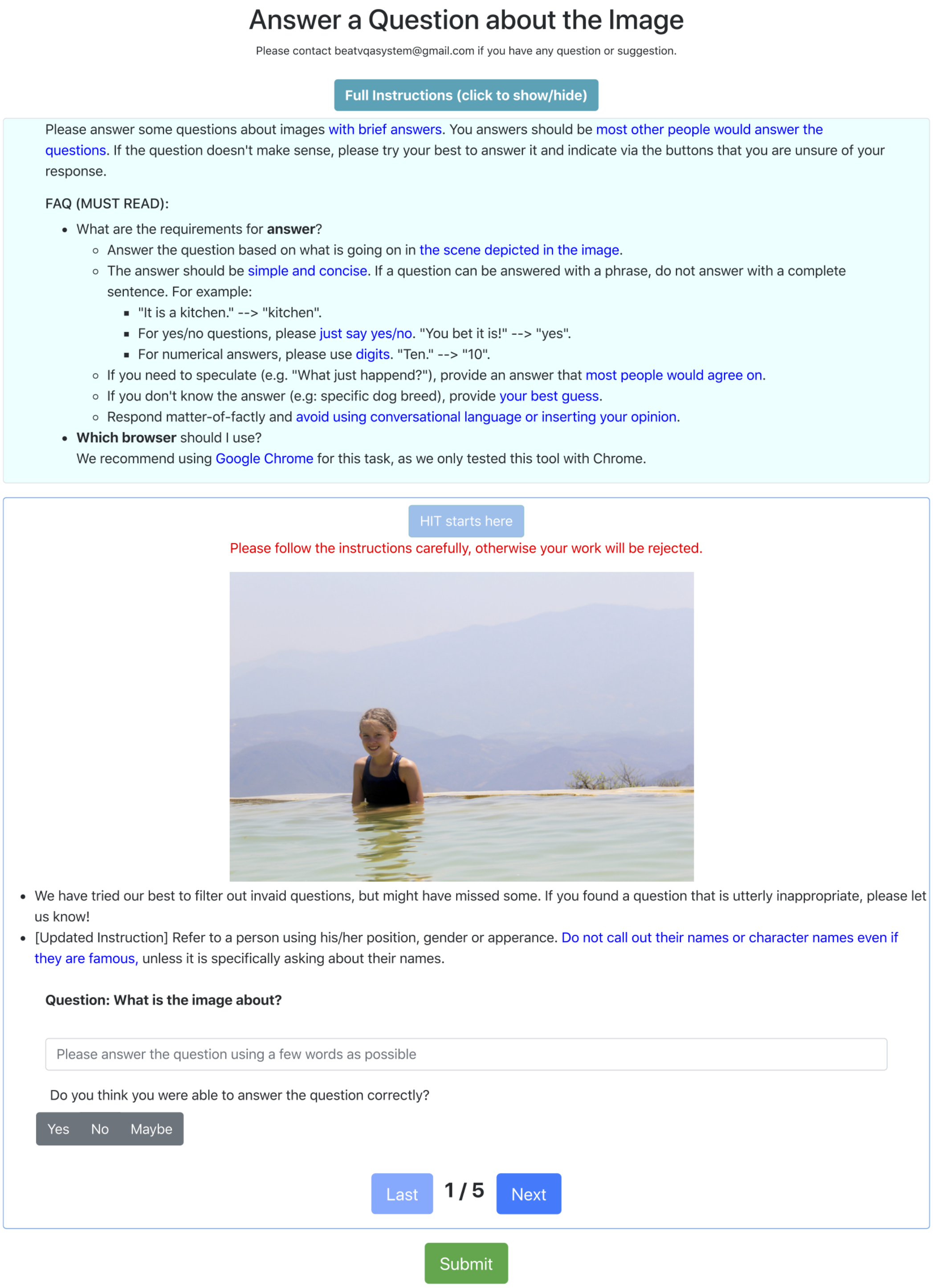
This box will show you all history of your attempts, including question, robot answer and your judgement.

#1 Question: ***Are there mountains in the background?*** Robot's answer: ***yes*** *with* confidence ***99.1%***. Your judgement: ***Definitely Correct***.

#2 Question: ***Is the little girl in an ocean?*** Robot's answer: ***yes*** *with* confidence ***78.2%***. Your judgement: ***Definitely Wrong***.

Figure 11: Example of model feedback shown to the annotators. After reviewing the model response, the annotator need to judge the correctness of the model answer ("Definitely Correct", "Not Sure", or "Definitely Wrong"). If the model answer is definitely wrong, the annotator is prompted to enter a correct answer.

ing. In *CVPR*, 2019. 3, 4



Answer a Question about the Image

Please contact beatvqasystem@gmail.com if you have any question or suggestion.

Please answer some questions about images with brief answers. You answers should be 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.

FAQ (MUST READ):

- What are the requirements for **answer**?
  - Answer the question based on what is going on in the scene depicted in the image.
  - The answer should be simple and concise. If a question can be answered with a phrase, do not answer with a complete sentence. For example:
    - "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 you need to speculate (e.g. "What just happend?"), 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.
- **Which browser** should I use?
  We recommend using Google Chrome for this task, as we only tested this tool with Chrome.

Please follow the instructions carefully, otherwise your work will be rejected.

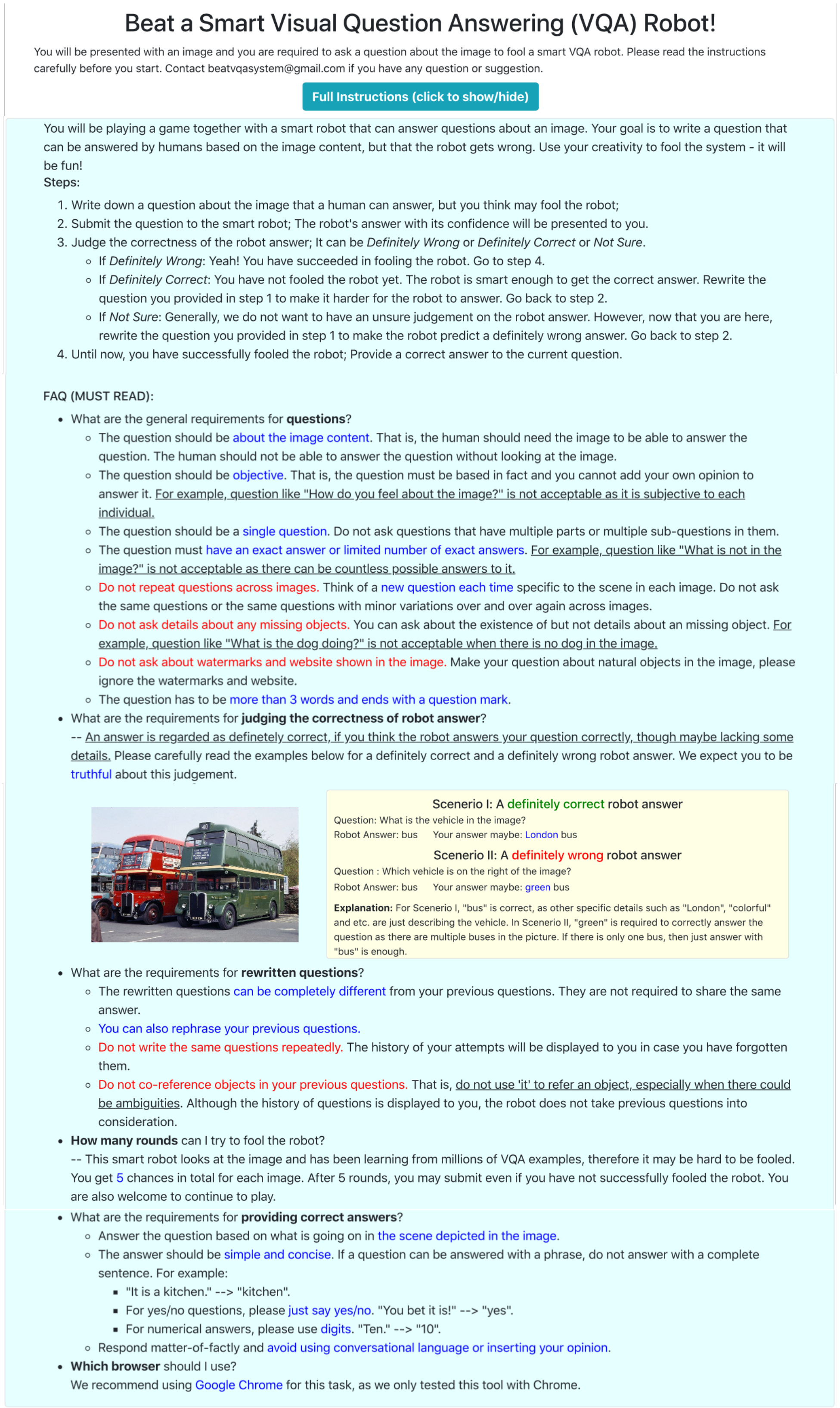
- We have tried our best to filter out invaid questions, but might have missed some. If you found a question that is utterly inappropriate, please let us know!
- [Updated Instruction] Refer to a person using his/her position, gender or apperance. Do not call out their names or character names even if they are famous, unless it is specifically asking about their names.

Question: What is the image about?

Please answer the question using a few words as possible

Do you think you were able to answer the question correctly?

Figure 12: UI for answer collection. Given an image and a question, an annotator is asked to write a concise answer to the question, and choose a confidence level for the answer ("Yes", "No", or "Maybe").



Beat a Smart Visual Question Answering (VQA) Robot!

You will be presented with an image and you are required to ask a question about the image to fool a smart VQA robot. Please read the instructions carefully before you start. Contact beatvqasystem@gmail.com if you have any question or suggestion.

You will be playing a game together with a smart robot that can answer questions about an image. Your goal is to write a question that can be answered by humans based on the image content, but that the robot gets wrong. Use your creativity to fool the system - it will be fun!

Steps:

1. Write down a question about the image that a human can answer, but you think may fool the robot;
2. Submit the question to the smart robot; The robot's answer with its confidence will be presented to you.
3. Judge the correctness of the robot answer; It can be *Definitely Wrong* or *Definitely Correct* or *Not Sure*.
   - If *Definitely Wrong*: Yeah! You have succeeded in fooling the robot. Go to step 4.
   - If *Definitely Correct*: You have not fooled the robot yet. The robot is smart enough to get the correct answer. Rewrite the question you provided in step 1 to make it harder for the robot to answer. Go back to step 2.
   - If *Not Sure*: Generally, we do not want to have an unsure judgement on the robot answer. However, now that you are here, rewrite the question you provided in step 1 to make the robot predict a definitely wrong answer. Go back to step 2.
4. Until now, you have successfully fooled the robot; Provide a correct answer to the current question.

FAQ (MUST READ):

- What are the general requirements for **questions**?
  - The question should be about the image content. That is, the human should need the image to be able to answer the question. The human should not be able to answer the question without looking at the image.
  - The question should be objective. That is, the question must be based in fact and you cannot add your own opinion to answer it. For example, question like "How do you feel about the image?" is not acceptable as it is subjective to each individual.
  - The question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
  - The question must have an exact answer or limited number of exact answers. For example, question like "What is not in the image?" is not acceptable as there can be countless possible answers to it.
  - Do not repeat questions across images. Think of a new question each time specific to the scene in each image. Do not ask the same questions or the same questions with minor variations over and over again across images.
  - Do not ask details about any missing objects. You can ask about the existence of but not details about an missing object. For example, question like "What is the dog doing?" is not acceptable when there is no dog in the image.
  - Do not ask about watermarks and website shown in the image. Make your question about natural objects in the image, please ignore the watermarks and website.
  - The question has to be more than 3 words and ends with a question mark.
- What are the requirements for **judging the correctness of robot answer**?
  -- An answer is regarded as definetely correct, if you think the robot answers your question correctly, though maybe lacking some details. Please carefully read the examples below for a definitely correct and a definitely wrong robot answer. We expect you to be truthful about this judgement.

Scenerio I: A definitely correct robot answer

Question: What is the vehicle in the image?

Robot Answer: bus   Your answer maybe: London bus

Scenerio II: A definitely wrong robot answer

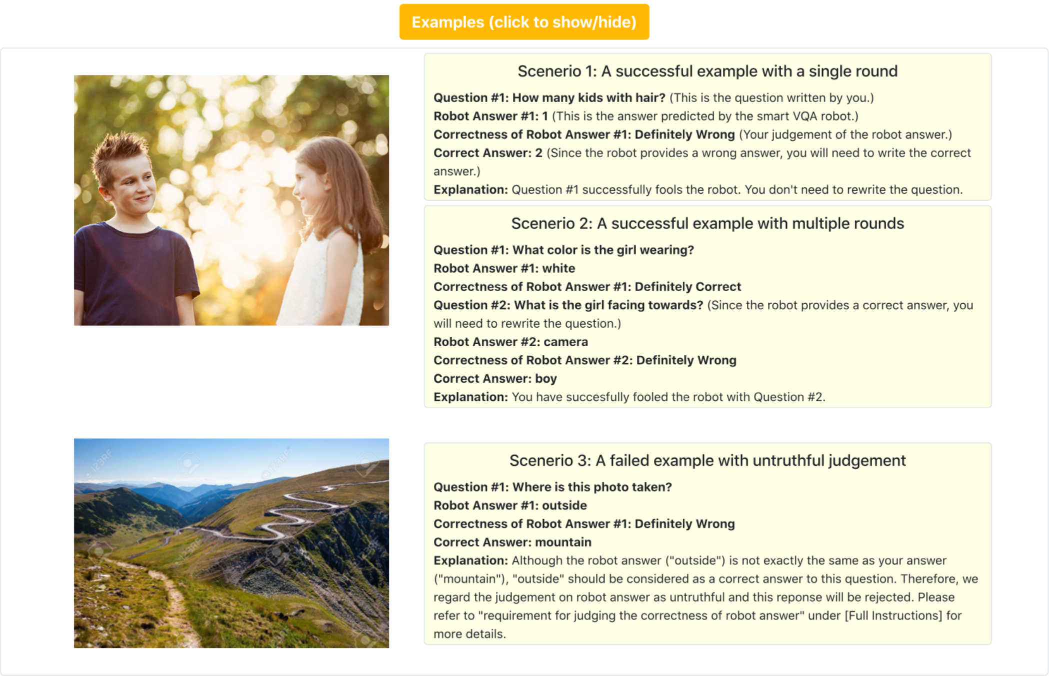
Question : Which vehicle is on the right of the image?

Robot Answer: bus   Your answer maybe: green bus

Explanation: For Scenerio I, "bus" is correct, as other specific details such as "London", "colorful" and etc. are just describing the vehicle. In Scenerio II, "green" is required to correctly answer the question as there are multiple buses in the picture. If there is only one bus, then just answer with "bus" is enough.

- What are the requirements for **rewritten questions**?
  - The rewritten questions can be completely different from your previous questions. They are not required to share the same answer.
  - You can also rephrase your previous questions.
  - Do not write the same questions repeatedly. The history of your attempts will be displayed to you in case you have forgotten them.
  - Do not co-reference objects in your previous questions. That is, do not use 'it' to refer an object, especially when there could be ambiguities. Although the history of questions is displayed to you, the robot does not take previous questions into consideration.
- **How many rounds** can I try to fool the robot?
  -- This smart robot looks at the image and has been learning from millions of VQA examples, therefore it may be hard to be fooled. You get 5 chances in total for each image. After 5 rounds, you may submit even if you have not successfully fooled the robot. You are also welcome to continue to play.
- What are the requirements for **providing correct answers**?
  - Answer the question based on what is going on in the scene depicted in the image.
  - The answer should be simple and concise. If a question can be answered with a phrase, do not answer with a complete sentence. For example:
    - "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".
  - Respond matter-of-factly and avoid using conversational language or inserting your opinion.
- **Which browser** should I use?
  We recommend using Google Chrome for this task, as we only tested this tool with Chrome.

Figure 13: Full instructions for question collection.



Scenerio 1: A successful example with a single round

Question #1: How many kids with hair? (This is the question written by you.)
Robot Answer #1: 1 (This is the answer predicted by the smart VQA robot.)
Correctness of Robot Answer #1: Definitely Wrong (Your judgement of the robot answer.)
Correct Answer: 2 (Since the robot provides a wrong answer, you will need to write the correct answer.)
Explanation: Question #1 successfully fools the robot. You don't need to rewrite the question.

Scenerio 2: A successful example with multiple rounds

Question #1: What color is the girl wearing?
Robot Answer #1: white
Correctness of Robot Answer #1: Definitely Correct
Question #2: What is the girl facing towards? (Since the robot provides a correct answer, you will need to rewrite the question.)
Robot Answer #2: camera
Correctness of Robot Answer #2: Definitely Wrong
Correct Answer: boy
Explanation: You have succesfully fooled the robot with Question #2.

Scenerio 3: A failed example with untruthful judgement

Question #1: Where is this photo taken?
Robot Answer #1: outside
Correctness of Robot Answer #1: Definitely Wrong
Correct Answer: mountain
Explanation: Although the robot answer ("outside") is not exactly the same as your answer ("mountain"), "outside" should be considered as a correct answer to this question. Therefore, we regard the judgement on robot answer as untruthful and this reponse will be rejected. Please refer to "requirement for judging the correctness of robot answer" under [Full Instructions] for more details.

Figure 14: Examples provided to annotators for question collection.