

A. Additional Dataset Analysis

In this section, we provide additional statistics and visualizations of the Visual7W dataset.

A.1. Question and Answer Lengths

We report the average question and answer lengths of each 7W category in Table 4, where the numbers in brackets show the standard deviations. The question and answer lengths are measured by their word counts. The *pointing* questions, which often describe some details of objects in the images, have longest average length (7.83 words). The *why* questions have the longest average answer length (3.21 words), as their answers often consist of phrases and sentences. Similarly, the *where*, *when* and *who* questions also have average answer lengths of more than two words. As mentioned in Sec. 4, the average question and answer lengths are 6.9 and 2.0 respectively.

Table 4: Average Question and Answer Lengths of 7W QA

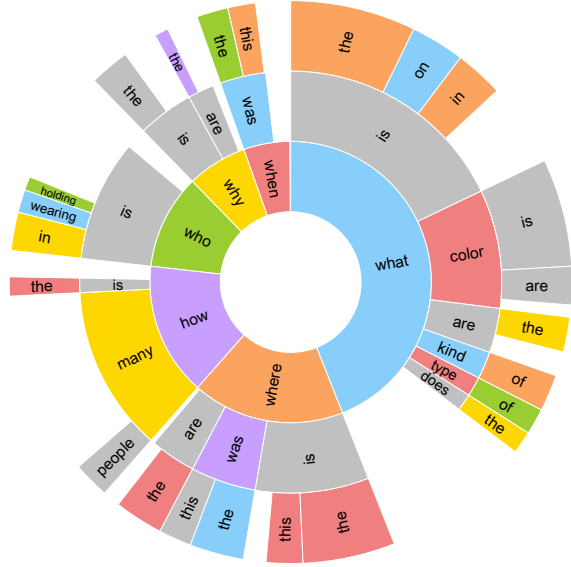
	What	Where	When	Who	Why	How	Which
Q	5.96 (1.8)	4.77 (1.0)	5.07 (1.1)	5.07 (1.3)	5.95 (1.5)	5.96 (1.6)	7.83 (2.4)
A	1.59 (1.1)	2.91 (1.3)	2.28 (1.4)	2.25 (1.4)	3.21 (1.8)	1.37 (1.1)	-

A.2. Question Distributions

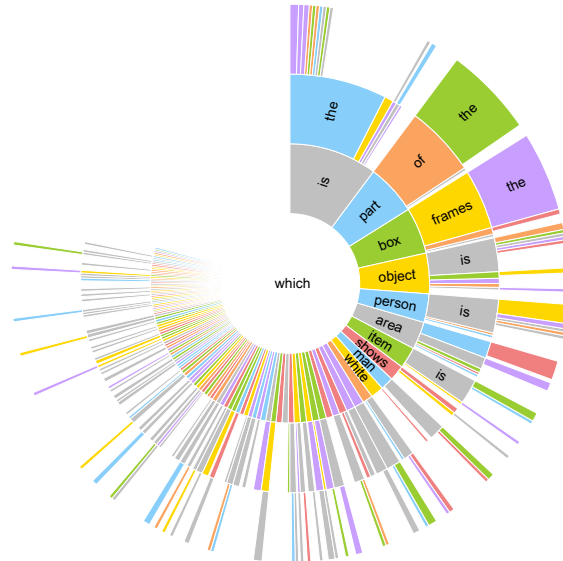
The questions can naturally be summed up into the 7W categories by their interrogative words. Within each category, the second and following words categorize the questions at increasing granularity. Inspired by VQA [1], we show the distributions of the questions by their starting words in Fig. 10. The *what* category is the most common category in *telling* QA. The other five categories make up for slightly more than half of all *telling* QA pairs. The starting words for the *pointing* (*which*) questions are scattered, as many *pointing* questions start with an object name.

A.3. Question and Answer Word Clouds

We visualize the word distributions of the questions and answers by word clouds. The font sizes in the clouds are proportional to relative frequency counts. Fig. 11 to Fig. 17 show the word clouds of questions and answers from each 7W category. The W words stand out as the largest word in each question cloud. The answer clouds show clear patterns of the 7W question categories: The *what* answers contain a cluster of words of colors and object names; however, the *where*, *when*, *who* and *how* answers contain many words of scene categories, time and dates, people and numbers (for *how many* questions) respectively. For *pointing* QA, the answers are object bounding boxes instead of sentences. Therefore, we visualize the object names of the boxes in the answer cloud of Fig. 17.



(a) Question distributions in *telling* QA



(b) Question distributions in *pointing* QA

Figure 10: Distribution of question types by starting words. The radians of the regions are proportional to the number of QA pairs within the corresponding categories.

B. Additional QA Examples

Fig. 18 shows additional examples of multiple-choice QA from the Visual7W dataset, in the same format as Fig. 2. These examples illustrate a diverse range of vision skills that are required for answering these questions. Among the examples of *telling* QA, question (c), (d) and (f) would require object classification, scene classification and action recognition; question (b) and (e) would require recogniz-



Figure 11: Word clouds of *what* questions and answers.

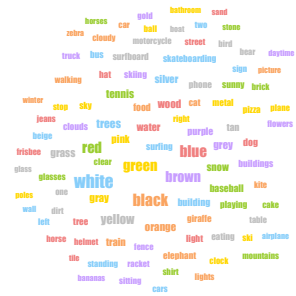


Figure 12: Word clouds of *where* questions and answers.



Figure 13: Word clouds of *when* questions and answers.



Figure 14: Word clouds of *who* questions and answers.



Figure 15: Word clouds of *why* questions and answers.

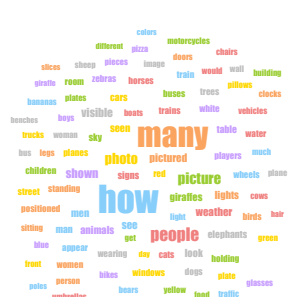


Figure 16: Word clouds of *how* questions and answers.



Figure 17: Word clouds of *which* questions and answers.

ing object attributes and fine-grained categories; beyond recognition-centered tasks, question (h), (i) and (j) would require high-level reasoning about space, events and common sense. Other *telling* QA examples are related to various vision tasks, such as counting (a), text detection (g), face recognition (k), and facial expression analysis (l). For the *pointing* QA examples, question (m), (n) and (p) would require spatial reasoning; question (o) and (r) would require reasoning about human action and object affordance respectively. In addition, all the questions are stated in natural language. It, therefore, requires joint reasoning between the textual and visual modalities to tackle the visual QA tasks.

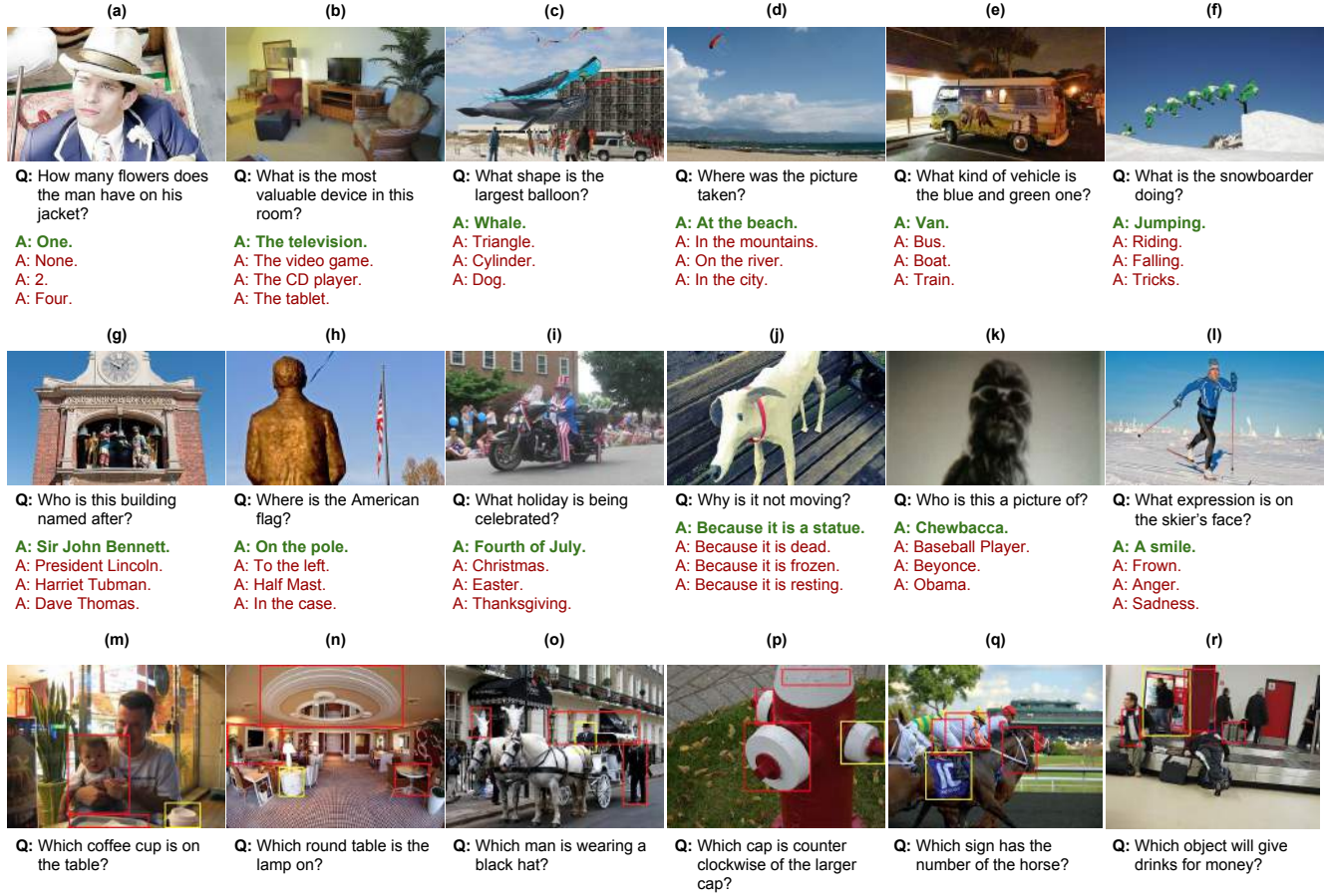


Figure 18: Additional QA examples from the Visual7W dataset. The first two rows show *telling* QA examples, where the ground-truth answers are shown in green, and three human-generated wrong answers are shown in red. The third row shows *pointing* QA examples, where the ground-truth answers are the yellow boxes, and three human-generated wrong answers are the red boxes.

Table 5: Model performances in the open-ended *telling* QA tasks (in top- k accuracy)

Method	What	Where	When	Who	Why	How	Overall
Top-10 Frequent Answers	0.133	0.000	0.358	0.000	0.000	0.403	0.140
Top-100 Frequent Answers	0.387	0.099	0.542	0.408	0.053	0.717	0.377
Top-500 Frequent Answers	0.588	0.381	0.671	0.566	0.154	0.782	0.557
Top-1000 Frequent Answers	0.660	0.464	0.731	0.665	0.259	0.802	0.628
Top-1 LSTM (Question)	0.174	0.045	0.243	0.161	0.083	0.223	0.156
Top-1 LSTM (Question + Image)	0.216	0.098	0.229	0.175	0.077	0.244	0.188
Top-5 LSTM (Question)	0.384	0.117	0.528	0.358	0.158	0.590	0.360
Top-5 LSTM (Question + Image)	0.432	0.262	0.546	0.398	0.156	0.594	0.413

C. Open-ended Telling QA Experiments

As mentioned in Sec. 6.1, an alternative method to evaluate the *telling* QA tasks is to let the model predict open-ended text outputs. Such freeform answers can be generated by the same attention-based QA model in Sec. 5, without re-training. Instead of picking answer candidates based

on perplexity, we use a different decoder, as used in the image captioning models [13, 44], to generate a new answer word from the previous LSTM hidden state and the last predicted answer word (or the question mark for the first answer word). This procedure continues until the period (i.e., the end token) is generated. One way to evaluate







					
Q: What is he inside?	Q: Where are the trees?	Q: When was this picture taken?	Q: Who is smaller?	Q: How many tomatoes are in the picture?	Q: Why is the man jumping?
A: Bath tub.	A: On the edge of the sidewalk.	A: During the day.	A: Bear on the left.	A: Four.	A: He is jumping to catch the frisbee.
w/o image T1: Food. T2: <empty> T3: Pizza T4: A surfboard T5: Water.	T1: On the ground. T2: <empty> T3: Background. T4: In the background. T5: In the sky.	T1: During the day. T2: Daytime. T3: At night. T4: Day time. T5: During daylight.	T1: The man. T2: A man. T3: Man. T4: The woman. T5: No one.	T1: Two. T2: One. T3: Three. T4: 2. T5: Four.	T1: To hit the ball. T2: To catch the ball. T3: Playing tennis. T4: He is playing tennis. T5: He is skiing.
w/ image T1: Toilet. T2: A toilet. T3: A mirror. T4: Bathroom. T5: A bathroom.	T1: On the sidewalk. T2: On the ground. T3: On the street. T4: On the road. T5: In the sky.	T1: During the day. T2: Daytime. T3: Day time. T4: During the daytime. T5: Daylight.	T1: The bear. T2: A bear. T3: Bear. T4: No one. T5: <empty>	T1: Two. T2: One. T3: Three. T4: Four. T5: 2.	T1: To hit the ball. T2: To catch the ball. T3: To catch the frisbee. T4: Playing. T5: To play

Figure 19: Qualitative results of attention-based QA model (see Sec. 5) on open-ended *telling* QA. T1 to T5 are the top-5 predictions in increasing order of perplexity. The model takes advantage of image content to make better predictions than without seeing the images. We mark the predictions that match exactly with the ground-truth answers in green. The model produces reasonable answers when seeing the images, although the generated answers in some cases fail to match with the ground-truth answers word by word.

open-ended answers is to match them with ground-truth answers word by word. This approach might work well for short answers, such as in DAQUAR [24], COCO-QA [33] and VQA [1]. However, it presents the *paraphrase* problem, especially for long answers: There might be multiple ways to paraphrase the correct answers. Previous work attempted to address this problem by using word ontology for short answers [24], or human judges for manual evaluation [7]. However, it remains an open question how to evaluate freeform answers in an accurate and automatic manner. To mitigate the paraphrase problem in open-ended evaluation, we let the model generate k different freeform answers. We use top- k accuracy to measure the performance, similar to the top-5 criterion used in ILSVRC [35]. We say a model is correct on a question if one of the k answers matches exactly with the ground-truth. We implement a beam search strategy [13] to make k different answer predictions from the model.

The first block in Table 5 shows the baseline performances of taking the top- k frequent answers in the training set as the predictions. These baseline results indicate the long-tail distribution of the answers, where the top-10, top-100, top-500 and top-1000 frequent answers only cover 14.0%, 37.7%, 55.7% and 62.8% of the answers in the test set. Among the *telling* QA categories, the *where* and *why* questions have the most diverse answers, where the top-1000 frequent answers cover only 46.4% and 25.9% of their answers respectively. This correlates with the answer length statistics in Sec. A.1, where the *where* and *why* questions have the longest average answer lengths among the *telling* QA categories.

The second block in Table 5 reports the results of

our model on top-1 and top-5 open-ended predictions. The model achieves a better overall accuracy with images (Question + Image) than without images (Question). However, the former performs slightly worse than the latter on both top-1 and top-5 predictions in the *why* categories. Overall, our model works better on the *what*, *when* and *how* categories, in comparison to the *where*, *who* and *why* categories, where the answers tend to be longer. The attention-based model achieves lower than 40% top-5 accuracy on the latter three categories.

We show qualitative examples of our model’s open-ended predictions in Fig. 19. The model produces better answers when the image features are fed into the model. For instance, in the first example of Fig. 19, the model predicts random objects as the answer when not seeing the image; whereas, it generates more relevant outputs, such as *toilet* and *bathroom*, when the image is shown to the model. Yet in many cases, the freeform outputs still fail to match with the ground-truth in the top-5 predictions, even though reasonable answers are generated. This illustrates the challenges of evaluating open-ended answers, especially in the presence of long answers. Therefore, we use multiple-choice tests as the default method to evaluate model performance in our QA tasks.