# A Additional details of dataset collection

## A.1 Examples of rejected questions

With a focus on overall question quality, we removed around 60% of questions written for having any of several flaws. The vast majority of questions removed exhibited one or more four flaws: 1) Only required recognition of a common object, 2) only required counting a readily specified object, 3) did not require looking at the image to answer, 4) only asked about the color of a readily specified object. Examples of questions from each of these categories are shown in Fig. 4.

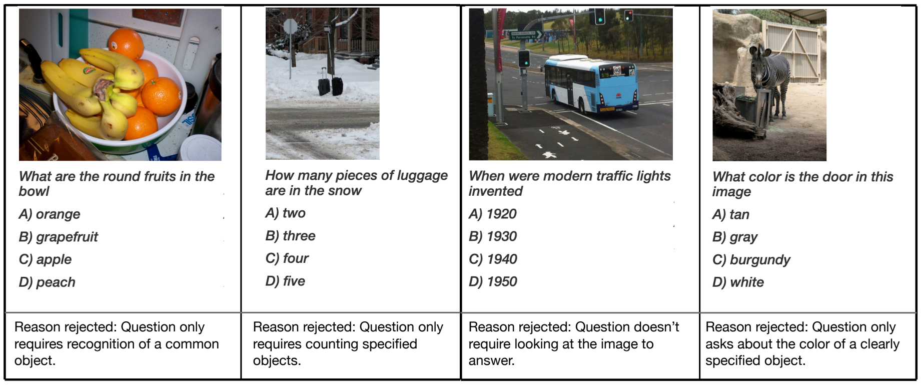
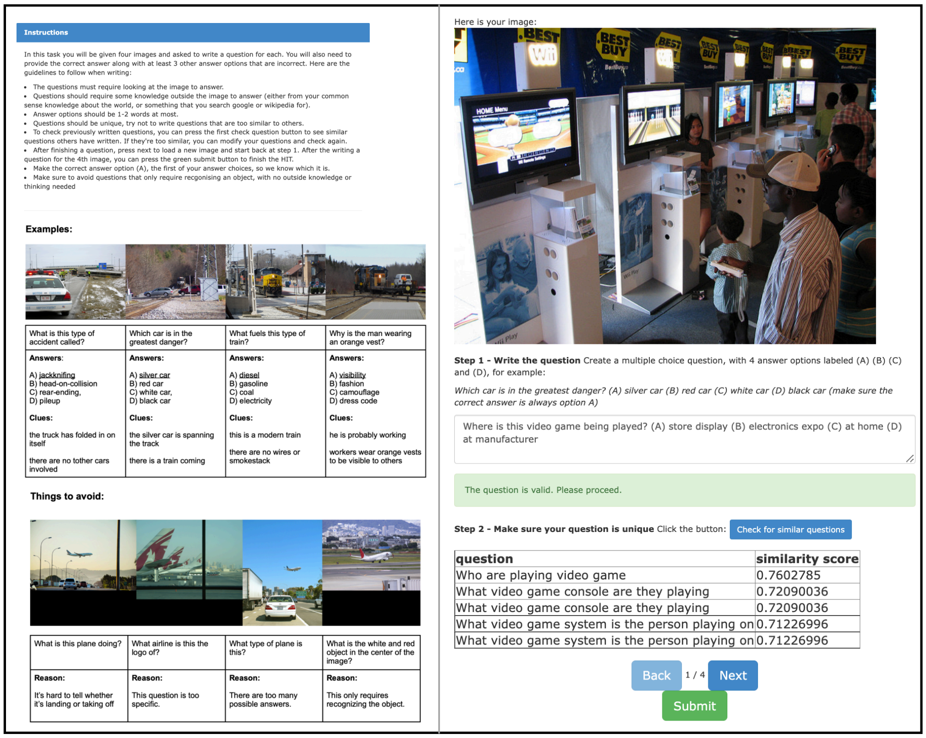

What are the round fruits in the bowl
A) orange
B) grapefruit
C) apple
D) peach
How many pieces of luggage are in the snow
A) two
B) three
C) four
D) five
When were modern traffic lights invented
A) 1920
B) 1930
C) 1940
D) 1950
What color is the door in this image
A) tan
B) gray
C) burgundy
D) white
Reason rejected: Question only requires recognition of a common object.
Reason rejected: Question only requires counting specified objects.
Reason rejected: Question doesn't require looking at the image to answer.
Reason rejected: Question only asks about the color of a clearly specified object.

Fig. 4: Examples of questions rejected for not meeting our criteria.

## A.2 Data collection interface

The data-collection interface used by crowdworkers to write questions is shown in Fig. 5. Detailed instructions along with examples of good and bad questions were provided. After writing a question, workers were required to press the "Check for similar question" button. This sent a request to a server which returned the five questions closest to those already written in our growing dataset. We asked workers to rewrite or rephrase questions that were too similar, but did not enforce a minimum distance cutoff. The set of questions queried were reset when collecting the val and test sets to allow a greater degree of overlap with the training set. After satisfied with their question, workers advanced to the next image. Each task workers performed included four images, nearby neighbors in a CLIP embedding space, which encouraged creative differences in questions written for similar images. Workers were only required to write two questions (out of four possible images) to allow them to skip images they didn't feel they could write a suitable questions for. This cut down on unsuitable questions that they would have otherwise been forced to write in order to complete the task.



**Instructions**

In this task you will be given four images and asked to write a question for each. You will also need to provide the correct answer along with at least 3 other answer options that are incorrect. Here are the guidelines to follow when writing:

- The questions must require looking at the image to answer.
- Questions should require some knowledge outside the image to answer (either from your common sense knowledge about the world, or something that you search google or wikipedia for).
- Answer options should be 1-2 words at most.
- Questions should be unique, try not to write questions that are too similar to others.
- To check previously written questions, you can press the first check question button to see similar questions others have written. If they're too similar, you can modify your questions and check again.
- After finishing a question, press next to load a new image and start back at step 1. After the writing a question for the 4th image, you can press the green submit button to finish the HIT.
- Make the correct answer option (A), the first of your answer choices, so we know which it is.
- Make sure to avoid questions that only require recgonising an object, with no outside knowledge or thinking needed

**Examples:**

| What is this type of accident called? | Which car is in the greatest danger? | What fuels this type of train? | Why is the man wearing an orange vest? |
|---|---|---|---|
| **Answers**:<br>A) jackknifing<br>B) head-on-collision<br>C) rear-ending,<br>D) pileup<br>**Clues:**<br>the truck has folded in on itself<br>there are no tother cars involved | **Answers:**<br>A) silver car<br>B) red car<br>C) white car,<br>D) black car<br>**Clues:**<br>the silver car is spanning the track<br>there is a train coming | **Answers:**<br>A) diesel<br>B) gasoline<br>C) coal<br>D) electricity<br>**Clues:**<br>this is a modern train<br>there are no wires or smokestack | **Answers:**<br>A) visibility<br>B) fashion<br>C) camouflage<br>D) dress code<br>**Clues:**<br>he is probably working<br>workers wear orange vests to be visible to others |

**Things to avoid:**

| What is this plane doing? | What airline is this the logo of? | What type of plane is this? | What is the white and red object in the center of the image? |
|---|---|---|---|
| **Reason:**<br>It's hard to tell whether it's landing or taking off | **Reason:**<br>This question is too specific. | **Reason:**<br>There are too many possible answers. | **Reason:**<br>This only requires recognizing the object. |

Here is your image:

**Step 1 - Write the question** Create a multiple choice question, with 4 answer options labeled (A) (B) (C) and (D), for example:

*Which car is in the greatest danger? (A) silver car (B) red car (C) white car (D) black car (make sure the correct answer is always option A)*

Where is this video game being played? (A) store display (B) electronics expo (C) at home (D) at manufacturer

The question is valid. Please proceed.

**Step 2 - Make sure your question is unique** Click the button: Check for similar questions

| question | similarity score |
|---|---|
| Who are playing video game | 0.7602785 |
| What video game console are they playing | 0.72090036 |
| What video game console are they playing | 0.72090036 |
| What video game system is the person playing on | 0.71226996 |
| What video game system is the person playing on | 0.71226996 |

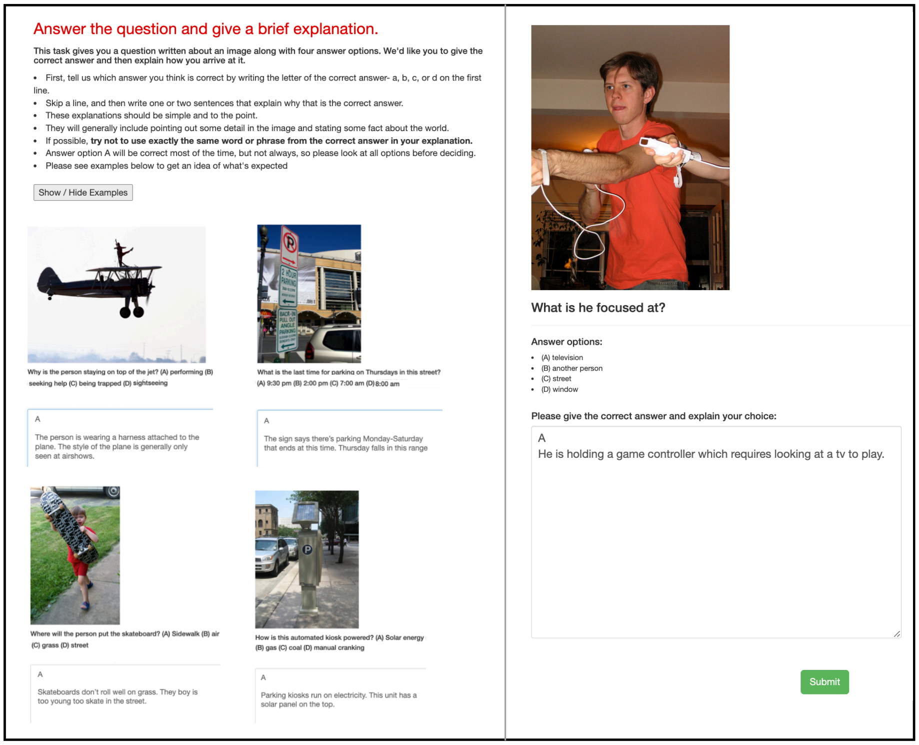
Fig. 5: Instructions and interface used for question collection.

After completing two questions, workers were allowed to submit their work and advance to the next image set.

The data-collection interface used by crowdworkers to write rationales is shown in Fig. 6. Detailed instructions along with examples of good rationales were provided. We first asked workers to confirm the correct answer or provide the answer they thought was correct. This allowed a check on the correctness of the original question, and questions with a disagreement were removed from the dataset. Workers then provided a 1-2 sentence explanation of why the answer was correct that included any external knowledge needed to arrive there.

## B Additional Details for Large-scale Pre-trained Models

We produce the vocabulary for the experiments in Sec. 5.2 from the training set by selecting all correct choices, as well as all choices and direct answers that appear in at least three questions. This results in a vocabulary with 10,424 answers.



Answer the question and give a brief explanation.

**This task gives you a question written about an image along with four answer options. We'd like you to give the correct answer and then explain how you arrive at it.**

- First, tell us which answer you think is correct by writing the letter of the correct answer- a, b, c, or d on the first line.
- Skip a line, and then write one or two sentences that explain why that is the correct answer.
- These explanations should be simple and to the point.
- They will generally include pointing out some detail in the image and stating some fact about the world.
- If possible, **try not to use exactly the same word or phrase from the correct answer in your explanation.**
- Answer option A will be correct most of the time, but not always, so please look at all options before deciding.
- Please see examples below to get an idea of what's expected

**Why is the person staying on top of the jet? (A) performing (B) seeking help (C) being trapped (D) sightseeing**

A

The person is wearing a harness attached to the plane. The style of the plane is generally only seen at airshows.

**What is the last time for parking on Thursdays in this street? (A) 9:30 pm (B) 2:00 pm (C) 7:00 am (D) 8:00 am**

A

The sign says there's parking Monday-Saturday that ends at this time. Thursday falls in this range

**Where will the person put the skateboard? (A) Sidewalk (B) air (C) grass (D) street**

A

Skateboards don't roll well on grass. They boy is too young too skate in the street.

**How is this automated kiosk powered? (A) Solar energy (B) gas (C) coal (D) manual cranking**

A

Parking kiosks run on electricity. This unit has a solar panel on the top.

**What is he focused at?**

**Answer options:**

- (A) television
- (B) another person
- (C) street
- (D) window

**Please give the correct answer and explain your choice:**

A
He is holding a game controller which requires looking at a tv to play.

Fig. 6: Instructions and interface used for rationale collection.

### B.1 Discriminative models

We train all of our discriminative models for 500 epochs with a learning rate of 0.01 and batch size of 128, except the model with ResNet input features, which is trained with a learning rate of 0.001.

### B.2 Contrastive models

The CLIP zero-shot setting requires no training. In the trained setting, we train our linear layer for 500 epochs with a learning rate of 0.01 and batch size of 128. We further elaborate on our "CLIP-style contrastive loss" below and visualize it in Fig. 7.

Recall that we have passed CLIP representations (for questions and/or images) through a linear layer to produce a 512-d embedding (the same size as a CLIP text encoding). For a batch of embeddings $E$ and the CLIP text encodings of their corresponding answers $A$, we produce a cosine similarity matrix between $E$ and $A$ (i.e. the purple matrix in Fig. 7, showing a batch size of 4). We apply softmax over each matrix row (producing embedding–answer matching proba-

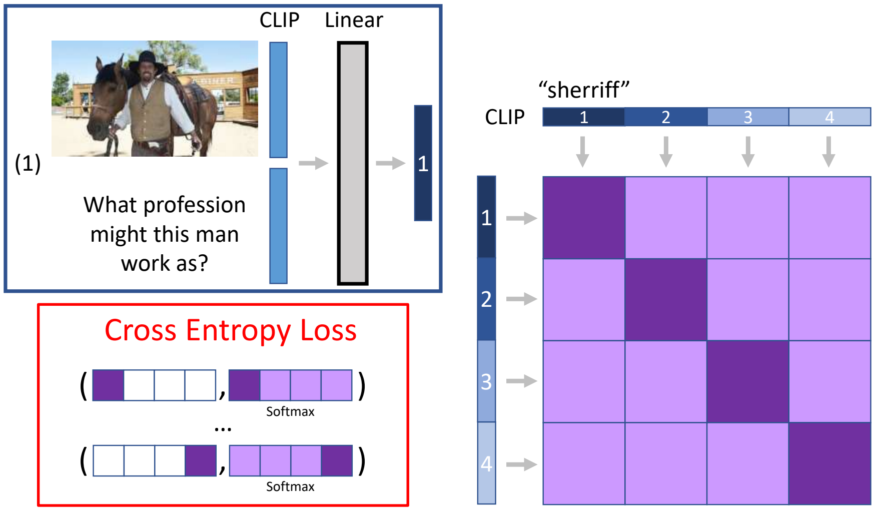


Fig. 7: As described in Sec. B.2. CLIP-style contrastive loss between embeddings (of questions and images) and CLIP text encodings (of answers). Shown for a batch size of 4.

bilities per embedding over answers in $A$) and compute a cross-entropy loss to maximize the similarity between each embedding and its corresponding answer.

### B.3 Generative models

We show our modified ClipCap model in Fig. 8. As in ClipCap [33], we provide CLIP image representations to a mapping network, which produces prefix tokens as input for GPT-2. We then tokenize our question and ground-truth answer (appended with an end-of-sequence string, ⟨EOS⟩) and also provide these tokens as input. The remaining input tokens (in black) are zero-padding. As mentioned in our paper, we also appended the (pre-tokenized) question string with "Choices: ..." during the MC setting.

This model is trained autoregressively. I.e., $O_i$ is generated conditionally, given $I_0 \cdots I_i$ (for input tokens $I$ and output logits $O$), and supervised with a cross-entropy loss against the next sequence token $I_{i+1}$. In our case, we only compute this cross-entropy loss for outputs corresponding with the ground-truth answer tokens (including ⟨EOS⟩).

At inference time, we prompt GPT-2 with our image prefix and question tokens. We have the model predict the most likely next token (i.e. generating a token in the answer) from the output logits. We append this token to the input and repeat this step, until the model predicts ⟨EOS⟩. We can use the tokenizer to decode these output tokens (excluding ⟨EOS⟩), producing our model's textual answer prediction. Note that beam search is an alternative way to generate text from autoregressive language models, but we found that it led to worse results, likely because the answers we are trying to generate are short (e.g. 1-3 words).

We fine-tuned the models in our experiments (choosing the checkpoint with the best F1 validation score for generated answers over 10 epochs), using the settings and COCO pre-trained weights (for the MLP mapping network) made available by the ClipCap authors [8]. For the pre-trained MLP model, they used CLIP ViT-B/32 features, produced 10 image prefix tokens, and had also fine-tuned GPT-2 (for their image captioning task). We further fine-tuned the GPT-2 weights on our task.

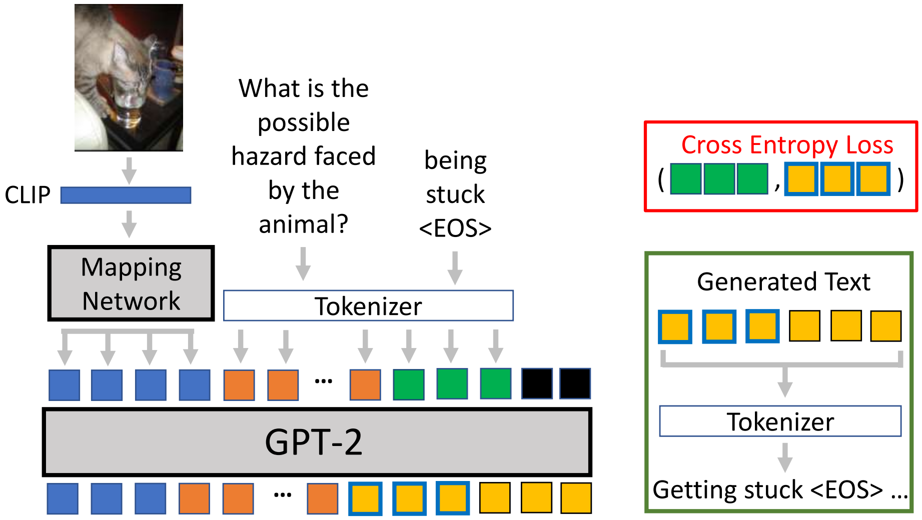


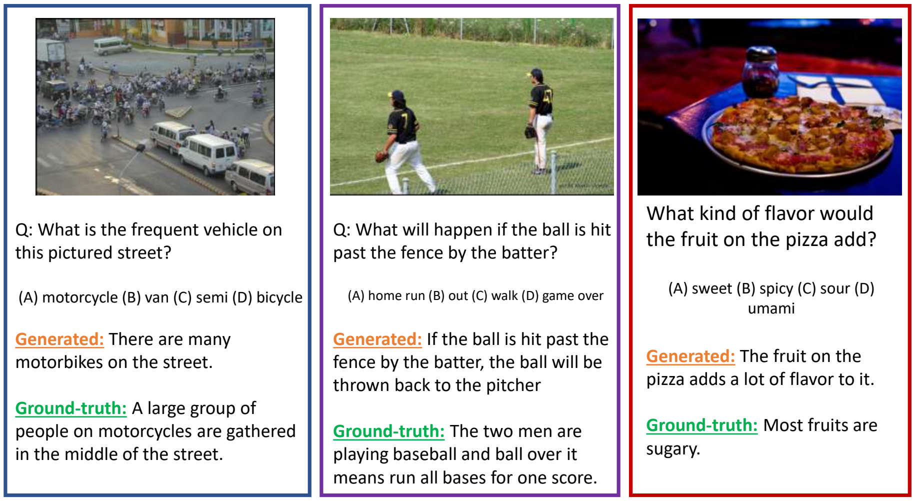
Fig. 8: Diagram of modified ClipCap architecture for VQA tasks.

## C Additional Details for Rationale Generation

We generated rationales from ClipCap in a nearly identical manner to how we generated answers (see Sec. B.3 and Fig. 8 above). However, we replace the ground-truth answer string/tokens with a ground-truth rationale. And, we don't provide "Choices: ..." in the ClipCap prompt for the MC setting. We also use beam search during generation, as it seems to perform better for these longer strings. We also use the MLP mapping network and continue to fine-tune GPT-2, as it demonstrates the best performance for this task. We again fine-tuned this model on our training data for 10 epochs and picked the checkpoints with best BLEU and METEOR validation scores.

We show some examples of generated rationales in Fig. 9.

[8] https://github.com/rmokady/CLIP_prefix_caption



Q: What is the frequent vehicle on this pictured street?

(A) motorcycle (B) van (C) semi (D) bicycle

Generated: There are many motorbikes on the street.

Ground-truth: A large group of people on motorcycles are gathered in the middle of the street.

Q: What will happen if the ball is hit past the fence by the batter?

(A) home run (B) out (C) walk (D) game over

Generated: If the ball is hit past the fence by the batter, the ball will be thrown back to the pitcher

Ground-truth: The two men are playing baseball and ball over it means run all bases for one score.

What kind of flavor would the fruit on the pizza add?

(A) sweet (B) spicy (C) sour (D) umami

Generated: The fruit on the pizza adds a lot of flavor to it.

Ground-truth: Most fruits are sugary.

Fig. 9: Examples of rationales generated by our modified ClipCap method for examples in our validation set.

## D Additional Details for Specialized Models

For all of these models, we use the same training hyperparameters as the original implementation. For all of the discriminative methods in the paper we use a fixed vocabulary constructed from direct answers that appeared two or more times in the training set. This includes 2,133 bi-grams or unigrams, with 1,937 words.

**Pythia** [20] Pythia is a modification of [1] that introduces changes to the architecture and learning schedule and utilizes more training data. We fine-tune it on the A-OKVQA dataset. For fine-tuning, we replace the top classification layer with a randomly initialized layer for our set of answer vocabulary.

**LXMERT** [45] LXMERT is a Transformer-based vision and language model pre-trained using a large amount of image-sentence pairs for a set of pre-training tasks such as masked language modeling and object prediction. The model is pre-trained on VQAv2 [12], GQA [14], VG-QA [56], COCO captions [7], and Visual Genome captions [23]. We then fine-tune the model using the training set of A-OKVQA.

**VilBERT** [28] ViLBERT is an extension of the BERT architecture to process vision and language modalities for learning a joint representation for them. ViLBERT has been pre-trained on proxy tasks, but it has been evaluated on VQA as a downstream task. ViLBERT is pre-trained using Conceptual Captions [42] and fine-tuned on A-OKVQA. To evaluate how well a model trained on VQAv2 or OK-VQA performs on A-OKVQA, we fine-tune ViLBERT after training them on thoese datasets. These models are referred to as 'ViLBERT-VQA' and 'ViLBERT-OK-VQA' in Table 5.

**KRISP** [**31**] KRISP is a method for knowledge-based VQA which combines multi-modal Transformers with graph neural networks methods on knowledge graphs. We use the same models and data and knowledge sources and pre-processing steps as in that work, but filter the knowledge graph based on A-OKVQA rather than OK-VQA (see Sec. 3.2 of [31]).

**GPV-2** [**22**] GPV-2 [22] is a generative vision and language model built using the T5 [39] language model and VinVL [54] image features. It was pre-trained on Conceptual Captions [42] and then fine-tuned in a multi-task setting on image captioning, visual question answering, object localization, and classification, as well on web-search images for 10,000 visual concepts.

We fine-tune the fully-trained model on A-OKVQA by training it to generate the most common answer for each question. For direct answer evaluations, answers are then generated using beam search with 20 beams. For multiple choice, the answers are ranked by the log-probability score assigned to them by the model.

We perform two additional experiments with rationales with this model. First, ground-truth rationales are appended to the question as additional input text. Recall that we do not provide rationales at test time. However, for this experiment we use them during test. We refer to this model as 'GPV-2 + GT Ratl.'. Second, we use the same setting, but we replace every occurrence of the ground-truth answer in the rationale with the [answer] token. We refer to this model as 'GPV-2 [22] + Masked Ans.' in Table 5.

## E Knowledge Type Results

We use the test subset that we collected knowledge types on (see Sec. 4) to look at the accuracy of these models for different types of knowledge. In Table 8, we see that while again GPV is the best overall and in every category, the results show some interesting distinctions. KRISP, which is specifically designed with access to explicit knowledge sources such as ConceptNet [26] performs better on "Knowledge Base" questions compared with other discriminative multi-modal transformer methods such as VilBERT and LXMERT as well compared to ClipCap which has an overall higher performance. It also performs better on "Physical Knowledge" which also tends to overlap with its knowledge sources.

Table 8: **Analysis of results based on knowledge type.**

| Model | Commonsense | Knowledge Base | Physical Knowledge | Visual Knowledge |
|---|---|---|---|---|
| VilBERT [28] | 24.30 | 19.96 | 29.76 | 26.55 |
| LXMERT [45] | 25.51 | 16.01 | 27.38 | 27.23 |
| KRISP [31] | 26.63 | 20.72 | 39.29 | 26.09 |
| ClipCap [33] | 27.19 | 16.57 | 30.95 | 33.41 |
| GR-GPT | 21.42 | 12.99 | 17.86 | 24.79 |
| GPV-2 [22] | 39.76 | 25.24 | 44.05 | 41.19 |

## F Biases

Our dataset has its own set of biases. Here are some examples: (1) The dataset is based on COCO, originally intended for identifying 80 object categories. Hence, the same biases exist in our dataset. For instance, because of the composition of those 80 categories, images of baseball fields and safaris (and hence questions) are more common than one might otherwise expect. (2) Selecting the choice that appears most frequently in the train set achieves above chance performance in the multiple-choice setting (although it performs poorly in the direct answer setting) as shown in Table 3-row (c). (3) For automated filtering, we used Pythia and RoBERTa trained on specific datasets. Hence, our data is biased by those methods as well. Regarding social biases, we checked the entire dataset and removed questions including offensive language, racial or gender biases, and stereotypes.