A Training Details

Except UNITER [11] and VILLA [17], we trained all models using MMF [55]. The evaluation results with the standard deviation over three different runs with three different seeds are provided in Table A.1. Below we detail the finetuning setup on the COCO [39] train 2017 split evaluated on the AdVQA validation split. Unless otherwise specified, we used the AdamW [33, 41] optimizer with an initial learning rate of $5 \times 10^{-5}$, epsilon of $1 \times 10^{-8}$, cosine schedule and a warmup of 2000 steps. All jobs are trained in distributed fashion using PyTorch [48] on NVIDIA V100 GPUs.

Table A.1: Model performance on VQA v2 and AdVQA val and test sets. * indicates that this model architecture (but not this model instance) was used in the data collection loop.

<table>
<thead>
<tr>
<th>Model</th>
<th>VQA test-dev</th>
<th>VQA test</th>
<th>AdVQA val</th>
<th>AdVQA test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human performance</td>
<td>80.78</td>
<td>91.18</td>
<td>84.73</td>
<td>87.53</td>
</tr>
<tr>
<td>Majority answer (overall)</td>
<td>-</td>
<td>13.38</td>
<td>24.67</td>
<td>11.65</td>
</tr>
<tr>
<td>Majority answer (per answer type)</td>
<td>-</td>
<td>27.59</td>
<td>31.01</td>
<td>29.24</td>
</tr>
<tr>
<td>Model in loop</td>
<td>MoViE+MCAN [45]</td>
<td>73.56</td>
<td>10.33</td>
<td>73.51</td>
</tr>
<tr>
<td>Unimodal</td>
<td>ResNet-152 [22]</td>
<td>26.37±0.38</td>
<td>10.85±0.37</td>
<td>24.82±0.27</td>
</tr>
<tr>
<td></td>
<td>BERT [14]</td>
<td>39.47±2.92</td>
<td>26.90±0.36</td>
<td>39.40±3.23</td>
</tr>
<tr>
<td>Multimodal (unimodal pretrain)</td>
<td>MoViE+MCAN* [45]</td>
<td>71.36±0.27</td>
<td>26.64±0.45</td>
<td>71.31±0.13</td>
</tr>
<tr>
<td></td>
<td>MMBT [30]</td>
<td>58.00±4.10</td>
<td>26.70±0.24</td>
<td>57.32±3.75</td>
</tr>
<tr>
<td></td>
<td>UniT [24]</td>
<td>64.34±0.05</td>
<td>32.12±0.21</td>
<td>64.32±0.04</td>
</tr>
<tr>
<td>Multimodal (multimodal pretrain)</td>
<td>VisualBERT [37]</td>
<td>70.37±0.05</td>
<td>28.70±0.36</td>
<td>70.05±0.11</td>
</tr>
<tr>
<td></td>
<td>ViLBERT [42]</td>
<td>69.42±0.30</td>
<td>27.36±0.18</td>
<td>69.27±0.14</td>
</tr>
<tr>
<td></td>
<td>ViLT [32]</td>
<td>64.52±0.42</td>
<td>27.11±0.14</td>
<td>65.43±2.43</td>
</tr>
<tr>
<td></td>
<td>UNITER Base [11]</td>
<td>71.87±0.01</td>
<td>25.16±0.49</td>
<td>70.50±0.08</td>
</tr>
<tr>
<td></td>
<td>UNITER Large [11]</td>
<td>73.57±0.21</td>
<td>26.94±0.31</td>
<td>72.71±0.22</td>
</tr>
<tr>
<td></td>
<td>VILLA Base [17]</td>
<td>70.94±1.25</td>
<td>25.14±0.93</td>
<td>69.50±1.44</td>
</tr>
<tr>
<td></td>
<td>VILLA Large [17]</td>
<td>72.29±0.39</td>
<td>25.79±0.46</td>
<td>71.40±0.37</td>
</tr>
<tr>
<td>Multimodal (unimodal pretrain + OCR)</td>
<td>M4C (TextVQA+STVQA) [25]</td>
<td>32.89±0.57</td>
<td>33.84±0.29</td>
<td>31.44±0.59</td>
</tr>
<tr>
<td></td>
<td>M4C (VQA v2 train set) [25]</td>
<td>67.66±0.34</td>
<td>36.57±0.36</td>
<td>66.21±0.38</td>
</tr>
</tbody>
</table>

MoViE+MCAN We use a batch size of 64 for 236K updates using a multi-step learning rate scheduler with steps at 180K and 216K, learning rate ratio of 0.2 and a warmup for 54K updates. Training takes an average of 2 days.

Unimodal We train the models with a batch size of 64 for 88K updates with linear learning rate schedule starting from $1 \times 10^{-5}$ with a warmup for 2000 updates. We used a linear learning rate schedule with 2000 warm up steps. The training takes an average of 8 hours.

MMBT [30] We trained MMBT from scratch with a batch size 64 without any pretraining following [30] for 88K updates. The training takes an average of 17 hours.

UniT [24] We initialized from the model pretrained on all 8 datasets [24] with COCO initialization. We set the batch size to 8, weight decay as $1 \times 10^{-4}$ and train the model on 8 GPUs for 2 days.

VisualBERT [37] We trained VisualBERT from the best pretrained model on COCO using MLM loss using a batch size of 64 which takes an average of 8 hours.

ViLBERT [42] We trained ViLBERT from the best pretrained model on Conceptual Captions using MLM loss using a batch size of 64 which takes an average of 13 hours.
ViLT [32] We trained ViLT with 44K updates, initial learning rate of $1e^{-4}$, eps as $1e^{-8}$ and weight decay as 0.01 for an average of 7 hours.

UNITER/VILLA [11, 17] We used the author-provided pretrained checkpoint for UNITER\(^3\) and VILLA\(^4\) with a confidence threshold of 0.075. The rest of the hyper parameters were consistent with the configuration provided in the repository. The training takes about 2 hours.

M4C [25] We used the same training schedule and hyper parameters as [25] used for TextVQA training [57] for both TextVQA + STVQA and VQA v2. We didn’t use extra Visual Genome [34] QA data due to lack of OCR text and used the OCR data from [25] for both COCO and TextVQA.

B Extended Dataset and Results Analysis

B.1 Question Types

To check what kind of questions types are present in AdVQA, we provide a detailed breakdown of question types based on categories similar to [4] in Table B.1.

Table B.1: **Detailed question type breakdown** for AdVQA’s full validation set along with percentage of those questions in “other”

<table>
<thead>
<tr>
<th>Question Type</th>
<th>full val set (%)</th>
<th>“others”</th>
</tr>
</thead>
<tbody>
<tr>
<td>what is</td>
<td>15.14</td>
<td>31.36</td>
</tr>
<tr>
<td>what color</td>
<td>5.57</td>
<td>12.96</td>
</tr>
<tr>
<td>what kind</td>
<td>0.85</td>
<td>1.98</td>
</tr>
<tr>
<td>what are</td>
<td>1.49</td>
<td>3.19</td>
</tr>
<tr>
<td>what type</td>
<td>0.85</td>
<td>1.98</td>
</tr>
<tr>
<td>is the</td>
<td>1.29</td>
<td>4.98</td>
</tr>
<tr>
<td>is this</td>
<td>1.02</td>
<td>0.35</td>
</tr>
<tr>
<td>how many</td>
<td>27.92</td>
<td>0.51</td>
</tr>
<tr>
<td>are</td>
<td>6.35</td>
<td>1.77</td>
</tr>
<tr>
<td>does</td>
<td>2.57</td>
<td>0.30</td>
</tr>
<tr>
<td>where</td>
<td>6.4</td>
<td>1.49</td>
</tr>
<tr>
<td>is there</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>why</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>which</td>
<td>3.26</td>
<td>7.40</td>
</tr>
<tr>
<td>do</td>
<td>0.67</td>
<td>0.02</td>
</tr>
<tr>
<td>what does</td>
<td>1.47</td>
<td>3.40</td>
</tr>
<tr>
<td>what time</td>
<td>0.76</td>
<td>1.77</td>
</tr>
<tr>
<td>who</td>
<td>0.36</td>
<td>0.84</td>
</tr>
<tr>
<td>what sport</td>
<td>0.12</td>
<td>0.28</td>
</tr>
<tr>
<td>what animal</td>
<td>0.32</td>
<td>0.74</td>
</tr>
<tr>
<td>what brand</td>
<td>0.93</td>
<td>2.16</td>
</tr>
</tbody>
</table>

B.2 Model Capabilities Required

To understand what capabilities are required for answering questions in AdVQA, we also sampled 50 random examples from the validation set. If more than one category is required for answering the questions, we include the question in all those categories. The categorization is based on the answer (i.e., what answering the question requires), rather than the question. Based on our analysis, we can see that AdVQA does require multiple kinds of reasoning capabilities in a model to solve it:

- 10 questions required color understanding.
- 5 questions’ answers were common objects (80 common object classes in COCO)
- 16 questions required reading or understanding text for answering
- 17 questions required positional understanding and reasoning.
- 5 questions required understanding uncommon objects
- 5 questions required external knowledge in some form.

\(^3\)UNITER Code: https://github.com/ChenRocks/UNITER
\(^4\)VILLA Code: https://github.com/zhegan27/VILLA
• At least 4 questions were directly solvable by common sense reasoning.
• One question required counting the number of the objects in the image.

B.3 BERT’s Performance

In comparison to multimodal models, BERT is very good at picking up on natural “biases” in AdVQA, which it probably learns from its pretraining on large scale text corpora. As seen in Table 4, the category-wise performance of BERT compared to multimodal is different as multimodal models perform better on numbers and others category while lagging on yes/no category. BERT’s performance is similar to majority class performance and majority class performance on all categories suggesting that BERT picks up on the dataset biases very well. We confirm this by doing an extensive analysis on BERT and VisualBERT’s predictions on AdVQA:

• 66.6% of the times BERT predicts “no” majority answer compared to 54.1% of the times for VisualBERT when the question is of “yes/no” type.
• For numeric questions, 45.6% times BERT predicts majority answer “2” whereas VisualBERT only predicts it 40.1% times.
• For questions requiring reading text (keywords “written”, “sign”, “text”), BERT mostly predicts “stop”.  
• For questions related to brand BERT mostly predicts “nike”, “coca-cola”, “ford” or “apple”.
• For countries, BERT mostly answers “usa”.
• For color, mostly predicts “black” or “white”.
• For questions asking what the person is holding predicts “tennis racket”.
• For cities, BERT mostly predicts “new york”.
• For position relation questions, BERT mostly predicts “right”.

B.4 Relationship to OKVQA, TextVQA and VQA-CP

In contrast to other VQA datasets which either target a very specific failure mode for VQA models or reorganize data distribution to reduce impact of bias, in AdVQA we create a more holistic and general dataset to make VQA more challenging. Specifically, compared to (i) OKVQA [44] in which questions specifically require external knowledge, (ii) TextVQA [57] requiring scene text reading and understanding, and (iii) VQA-CP [2] specifically changing answer distribution between train and test splits, we don’t ask annotators to specifically target a fixed failure mode or artificially generate an unbiased question distribution but take a more holistic model-in-the-loop approach. In our opinion, asking annotators to fool an existing VQA model covering all of these failure modes and biases (and more) makes AdVQA more suitable to be the next generation of a generic VQA benchmark. Having these alternative benchmarks has helped certainly push progress and put a magnifying lens on very difficult failure modes individually and we consider them complementary to AdVQA.

C Annotation details

Before annotators were able to proceed with the main task, they had to pass an on-boarding phase (Section C.1). We restricted the interface to be only accessible on non-mobile devices, to annotators in the US with at least 100 approved hits and with an approval rate higher than 98%. The annotators were paid a bonus if their self-claimed fooling question was verified to be fooling and valid by two other annotators.

C.1 Qualification Phase

The annotators were asked to go through a two-stage qualification phase. In the first stage, they were shown 11 examples that include both valid and invalid questions. Figure B.1 shows valid and invalid examples in the Example Stage. After scrolling through those 11 examples, the annotators proceeded to the second stage, in which we ask them to complete a quiz. The annotators passed only if they got more than 6 out of 7 correct, after which they qualified for the main task. There are two types of quiz questions: 1) determine if the provided answer is correct for the specific image question pair; and 2) determine if a given question is valid with the image as context. See Figure B.2 for examples of those
Examples

(a) Instructions for Examples Stage

In this section please carefully go through each example of valid and invalid questions.

A question is considered invalid if any of the following conditions is met:

- The question does not require the image to answer.
  i.e., What is the capital of the USA?
- The answer is not commonly known to other people.
  i.e., “What is the name of this plant?” when very few people would be able to recognize the plant and know the name.
- The question is not based on the scene depicted in the image or the answer could not be provided correctly based on the image.
  i.e., “What is the brand of the soap?” when the brand name of the soap is only partially visible from the image.
- “What is the woman doing?” when there is no woman in the image.

For each example below, we show an image, a question and a description of why the question is valid or invalid. Additionally, for the valid questions we include the answer the AI produced, and whether the answer was correct or not.

Please go through each example carefully.

(b) Invalid example

Figure B.1: Qualification Phase Stage 1: View Examples

two question types. If an annotator failed the first time, they were given an explanation on the correct choice before being allowed a second try on a different (but similar) set of questions.

C.2 Main Labeling Task

Figure B.3a and B.3b show the instructions given to first-time annotators for the “question collection” and “question validation” tasks. The instructions were hidden for the non-first-time annotators, but remained accessible via a button at the top. Figure B.4a and B.4b show the preview landing pages on Mechanical Turk.

C.3 Answer collection task

Figure C.1 and C.2 show the preview and main interface of the answer collection stage. For each question, we collect ten answers from ten different annotators using this interface. We provide explicit instructions following [20] and [57] to avoid ambiguity and collect short relevant answers.

The annotators are also provided a checkbox to select “unanswerable” in case the question is ambiguous or can’t be answered which annotators are suggested to use sparingly. Finally, we use and show a set of hand-crafted already annotated questions without ambiguity randomly to filter out bad annotators by comparing their answers to ground truth. An annotator is prevented from doing the task if they fail the test three times.
Quiz - Judge whether a question is valid, and judge the AI’s answer

Now that you have seen examples, let’s test your understanding! The first part of this test consists of **validating the AI’s answers**. For every image, question and corresponding AI answer below, tell us if the AI answer was correct or not. The second part consists of **validating questions** that the AI receives. For every question tell us if it is valid or not. Remember:

A question is considered **invalid** if any of the following conditions is met:

- The question does not require the image to answer.
  - i.e., What is the capital of the USA?
- The answer is not commonly known to other people.
  - i.e., “What is the name of this plant?” when very few people would be able to recognize the plant and know the name.
- The question is not based on the scene depicted in the image or the answer could not be provided correctly based on the image.
  - i.e., “What is the brand of the soap?” when the brand name of the soap is only partially visible from the image.
- i.e., “What is the woman doing?” when there is no woman in the image.

**WARNING:** You have to correctly validate at least 6 out of the 7 examples to unlock the main task. If you fail you will not be able to complete the onboarding.

**WARNING:** If you do not follow our instructions (shown above), your work will be rejected and you will be banned.

---

Figure B.2: Qualification Phase Stage 2 - quiz. Annotators are allowed access to the main task only if they passed the quiz.

---

(a) Instructions for Quiz Stage

- **Question Collection Instructions**
  - Ask questions and host the AI
  - In the box, you will learn to ask questions about an image that test if you know the answering hierarchy. The AI is reasonably good at understanding English pronunciation.
  - Given an image, you are presented in the following:
    - A right answer found in the image is a valid question has the following characteristics:
      - The question requires the image to answer.
      - The image contains an object that is clearly visible.
      - The object is not easily recognizable from the scene.
    - An example is “The book on the table,” with the image of a book on a table.
    - Another example is “The cat is sitting on the floor,” with the image of a cat sitting on the floor.
    - A question is invalid if it does not require the image to answer.
      - i.e., “What is the capital of the USA?”
  - **Validating the AI answer**
    - First, answer the question with your best guess.
    - Then, check the correct answer provided by the AI.
    - If the AI answer is correct, click “Correct.”
    - If the AI answer is incorrect, click “Wrong.”

(b) Determine if the answer is correct

(c) Determine if the question is valid

---

Figure B.3: Main Labeling Task Instructions

**Validation Examples**

- Front view of a cat in a box: The task consists of two parts. First, you have to determine if the question is valid. Most of the questions involve the cat, so you should check whether the cat is visible in the image. If the task seems to have an answer, you have to answer the question. If the task does not have an answer, you have to answer it with the word “uncertain.”
- “What is the capital of the USA?”
- “What is the name of this plant?”
- “What is the brand of the soap?”
- “What is the woman doing?”

**Question Collection Instructions**

- **Task Collection**
  - Ask the question on an image.
  - The question requires the image to answer.
  - The image contains an object that is clearly visible.
  - The object is not easily recognizable from the scene.
- **Validating the AI answer**
  - First, answer the question with your best guess.
  - Then, check the correct answer provided by the AI.
  - If the AI answer is correct, click “Correct.”
  - If the AI answer is incorrect, click “Wrong.”

---

Figure B.4: Example Image: a cat in a box.
Figure B.4: Preview - landing page on MTurk interface.

Figure B.5: Validation Interface

WARNING: if you do not follow our instructions (shown above), your work will be rejected and you will be banned.
Answer 10 questions on images

Description

Please answer some questions about images with text answers. Your answers should be how most other people would answer the question. If the question doesn’t make sense, please try your best to answer it and indicate with a blank cell that you are unsure of your response.

If you don’t follow the following instructions, you will be disqualified from this task.

- Answer the question based on what you see in the image, not what you don’t see.
- Your answer should be one word or a complete sentence.
- "Yes" or "No" answers are fine. If you don’t know, say "I don’t know".
- If you say "I don’t know", you may say why.
- "Yes" or "No" answers, please put yes.
- "No" or "False", please put false.
- All numbered answers, please put digit.
- "No" or "False", please put 0.
- If you uncertainly answer (e.g., what just happened?), provide an answer that most people would agree on.
- Give your answer. For example, "I don’t know". Please answer your best.
- Provide your answer with the most appropriate choice of words or phrases.
- Repeat the word or phrase and don’t use alternative words or phrases.
- If you feel that the question is not clear, please describe your answer by your question. If you have any questions, please ask before starting.

Please read instructions carefully so that you don’t get disqualified. You will only be disqualified if you are not paying proper attention.

We have provided some functionality on the image to help with readability and clarity. If necessary, please use your browser’s zoom to zoom into the image to see the image more clearly. You can also “right-click” on the image and click “Open Image in new tab”.

After entering the answer, click “Close with HIT” button to complete.

Completion

To complete the task, you must enter an answer to the question in the text box and then click “Submit”.

Reward

On successfully completing the task, you will receive a reward of $3.00.

Close Window/Pause/Return HIT

Once the task has started, close window/Pause or return HIT will result in HIT EXPired if you did not complete the task.

Important Notice

We generally review HITs and your work will get reviewed if you do not follow the instructions above.

Also, note that on very rare occasions, images can be disturbing in nature. Please keep that in mind while you work on this task. If you are ready, please click “accept” to start the task.

Figure C.1: Preview for Answer Collection

Figure C.2: Answer collection interface
D  Random Samples

We show 10 randomly selected samples in Figure D.1.

### Table D.1: Random examples from AdVQA.

<table>
<thead>
<tr>
<th>Image</th>
<th>AdVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image.png" alt="Image" /></td>
<td><img src="image.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Q:** Which hands does he have bracelets on?
**Processed Answers:** both (count: 9)
**Raw Answers:** 'both', 'both hands', 'both', 'both', 'both', 'both', 'both', 'both', 'both', 'both'

**Q:** What is the baby wearing?
**Processed Answers:** overalls (count: 6)
**Raw Answers:** 'shirt & overall', 'overalls', 'overalls', 'overalls', 'jumper', 'overalls', 'coverals', 'dungerees', 'overalls', 'overalls'
Q: How many people can be seen in the room?
Processed Answers: 6 (count: 5), 7 (count: 3)

Q: What is on the stovetop?
Processed Answers: kettle (count: 4)
 ’tea kettle’, ’kettle’, ’teapot’, ’kettle’, ’kettle’

Q: What does the three letters spell?
Processed Answers: unknown (not in vocab)
Raw Answers: ’pub’, ’pub’, ’pub’, ’pub’
 ’pub’, ’pub’, ’pub’, ’pub’, ’pub’

Q: Are the windows all the same size?
Processed Answers: no (count: 10)
<table>
<thead>
<tr>
<th>Image</th>
<th>AdVQA</th>
</tr>
</thead>
</table>
| ![Image](image.jpg) | **Q**: how many pieces of meat?  
**Processed Answers**: 2 (count: 7), 1 (count: 3)  
**Raw Answers**: '2', '2', '2', '2', '2', '2', '2', '1', '1', '1', '1', '2' |
| ![Image](image.jpg) | **Q**: Which is the largest window?  
**Processed Answers**: middle (count: 3)  
**Raw Answers**: 'center', 'bottom middle', 'bottom middle', 'middle', 'middle', 'bottom middle', 'middle lower', 'where a cat sits', 'middle one', 'the middle' |
| ![Image](image.jpg) | **Q**: What color is the checkerboard background?  
**Processed Answers**: black and white (count: 5)  
**Raw Answers**: 'unanswerable', 'black and white', 'gray', 'black and white', 'black', 'black white', 'black and white', 'black and white', 'black white', 'black and white' |
| ![Image](image.jpg) | **Q**: What does it say on the side of the boat closest in the foreground?  
**Processed Answers**: unknown (not in vocab)  
**Raw Answers**: 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci', 'sanssouci' |