Supplementary Material of “From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models”

Anonymous CVPR submission

Paper ID 12114

A. Appendix

A.1. Reproducibility Statement

We acknowledge the importance of reproducibility for research work and try whatever we can to ensure the reproducibility of our work. As for the implementation of our method, details such as hyperparameters are provided in Section 4.1 in the main paper. We will publicly release all codes after the acceptance of this paper.

A.2. Broader Impact Statement

We acknowledge that while the Img2LLM achieves comparable or superior performance to other zero-shot VQA methods, it has not reduced the inherent bias of these systems. Social-economic biases based on gender, age, race, and ethnicity exist in the datasets, LLMs, and VQA systems presented in this paper, including Img2LLM. Future work could assess the magnitude of this bias and mitigate its impact.

A.3. Details about Question-Relevant Caption Generation

Concretely, we denote features of image patches extracted by ITE as $f^i_v \in \mathbb{R}^{K \times D^i_v}$ and question features as $f^i_q \in \mathbb{R}^{L \times D^i_q}$, where $i$ is the number of the layer of ITE, $K$ is the number of images patches, $L$ is the number of token in the given question, $D^i_v$ is the dimension of patch feature in the $i$-th layer of ITE network and $D^i_q$ is the dimension of textual feature in the $i$-th layer of ITE network. For cross-attention head in $i$-th layer, the cross-attention scores $W^i$ between each image patch and each token in question can be calculated directly as

$$W^i = \text{softmax} \left( \frac{f^i_v W^i_q W^i_k ^\top f^i_v ^\top}{\sqrt{D^i_q}} \right).$$  \hspace{1cm} (1)

where $W^i_q \in \mathbb{R}^{D^i_q \times D^i_q}$ is the query head and $W^i_k \in \mathbb{R}^{D^i_k \times D^i_q}$ is the key head in the $i$-th layer of ITE network. With Equation 1, we obtain a cross-attention matrix $W^i \in \mathbb{R}^{L \times K}$, where each row is the cross-attention scores of each token in the question over all image patches. Specifically, the attention matrix $W^i$ can be regarded as the patch importance for ITE to calculate the similarity of whole image and question, but it still contains redundancy that contributes only a minor performance loss [1], indicating that some patches are uninformative. In order to find these less relevant image patches, we follow GradCAM and compute the derivative of the cross-attention score from ITE function $\text{sim}(v, q)$, i.e., $\partial \text{sim}(v, q) / \partial W$, and multiplying its gradient matrix with the cross-attention scores element-wisely. The relevance of the $k$-th image patch with the question, $r^i_k$, can be computed as the average over $H$ attention heads and the sum over $L$ textual tokens:

$$r^i_k = \frac{1}{H} \sum_{h=1}^H \min \left( 0, \frac{\partial \text{sim}(v, q)}{\partial W^i} \right) W^i_{kh},$$  \hspace{1cm} (2)

where $h$ is the index of attention heads and $i$ is the layer index of ITE.

A.4. Experimental Results of Supervised Learning Methods in A-OKVQA

We show the experimental comparisons between our method and supervised model on A-OKVQA dataset [6] as Table 3 shows. We can observe that our method outperform almost all supervised model with smaller size language model. This strongly support our method’s effectiveness in leveraging reasoning power of large language models.

A.5. Template-Based Question Design

We design question templates for each part of speech type of answers as Table 2 shows.

A.6. Sensitive Analysis

We evaluate the sensitive analysis about the QA pairs and number of captions in prompt for LLM as Table 3 shows. We can observe that the differences in QA scores on OKVQA dataset are not higher than 1 as long as QA pairs in prompts. The results demonstrate the performance of our
Table 1. The experimental comparisons with models trained in A-OKVQA training dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>A-OKVQA</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Models Fine-Tuned in A-OKVQA Training Set</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViLBERT [4]</td>
<td>30.6</td>
<td>25.9</td>
<td></td>
</tr>
<tr>
<td>LXMERT [7]</td>
<td>30.7</td>
<td>25.9</td>
<td></td>
</tr>
<tr>
<td>KRISP [5]</td>
<td>33.7</td>
<td>27.1</td>
<td></td>
</tr>
<tr>
<td>GPV-2 [3]</td>
<td><strong>48.6</strong></td>
<td></td>
<td><strong>40.7</strong></td>
</tr>
<tr>
<td><strong>Zero-Shot Evaluation with Plug-in Frozen Large Language Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>33.3</td>
<td>32.2</td>
<td></td>
</tr>
<tr>
<td>Ours_{11B}</td>
<td>33.3</td>
<td>33.0</td>
<td></td>
</tr>
<tr>
<td>Ours_{30B}</td>
<td>36.9</td>
<td>36.0</td>
<td></td>
</tr>
<tr>
<td>Ours_{668B}</td>
<td>38.7</td>
<td>38.2</td>
<td></td>
</tr>
<tr>
<td>Ours_{175B}</td>
<td>42.9</td>
<td></td>
<td><strong>40.7</strong></td>
</tr>
</tbody>
</table>

Table 2. The question templates for answers with different part of speech.

<table>
<thead>
<tr>
<th>Part of Speech of Answer</th>
<th>Question Templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>What item is this in this picture?</td>
</tr>
<tr>
<td></td>
<td>What item is that in this picture?</td>
</tr>
<tr>
<td>Verb</td>
<td>What action is being done in this picture?</td>
</tr>
<tr>
<td></td>
<td>Why is this item doing in this picture?</td>
</tr>
<tr>
<td></td>
<td>Which action is being taken in this picture?</td>
</tr>
<tr>
<td></td>
<td>What action is item doing in this picture?</td>
</tr>
<tr>
<td></td>
<td>What action is item performing in this picture?</td>
</tr>
<tr>
<td>Adjective</td>
<td>How to describe one item in this picture?</td>
</tr>
<tr>
<td></td>
<td>What is item's ADJ TYPE in this picture?</td>
</tr>
<tr>
<td></td>
<td>What is the ADJ TYPE in this picture?</td>
</tr>
<tr>
<td>Num</td>
<td>How many things in this picture?</td>
</tr>
</tbody>
</table>

Table 3. The experimental results of using different number of captions and QA pairs as prompts. The experiments are run on OK-VQA with OPT 30B.

<table>
<thead>
<tr>
<th>QA Pairs</th>
<th>Caption</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.3</td>
<td>19.6</td>
<td>22.7</td>
<td>23.4</td>
<td>24.0</td>
<td>24.8</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>40.9</td>
<td>41.6</td>
<td>42.1</td>
<td>42.1</td>
<td>41.9</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>41.2</td>
<td>41.3</td>
<td>41.3</td>
<td>41.7</td>
<td>42.2</td>
<td>42.0</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>41.0</td>
<td>41.0</td>
<td>41.7</td>
<td>41.8</td>
<td>41.6</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>40.3</td>
<td>40.7</td>
<td>40.6</td>
<td>40.3</td>
<td>40.3</td>
<td>41.1</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>40.6</td>
<td>40.6</td>
<td>40.7</td>
<td>40.9</td>
<td>40.6</td>
<td>41.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. The experimental results of using different number of patches to generate question-relevant captions. The experiments are run on OK-VQA with OPT 30B.

<table>
<thead>
<tr>
<th>Patch_num</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>41.2</td>
<td>41.8</td>
<td>41.6</td>
<td>39.8</td>
</tr>
</tbody>
</table>

Table 5. The experimental results of generating different number of question-relevant captions. The experiments are run on OK-VQA with OPT 30B.

<table>
<thead>
<tr>
<th>Caption_num</th>
<th>PICA</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17.7</td>
<td>38.3</td>
<td>40.9</td>
<td>41.4</td>
<td>41.8</td>
</tr>
</tbody>
</table>

A.7. Examples
Figure 1. Success case analysis for OK-VQA. Green color indicates answer cues and correct prediction.
Question: why is timing of the essence when delivering this food item? **GT answer:** temperature/hot still/stay ... large pizzas sitting next to each other? **Answer:** boxes
**Synthetic question 2:** where are two large pizzas sitting next to each other? **Answer:** table
**Question:** why is timing of the essence when delivering this food item? **Predicted answer:** hot

**Question:** what era is this furniture from? **GT answer:** victorian/1940s
**Caption 1:** a living room with a small television in front of the window
**Caption 2:** a vintage tv is sitting on a nice table in the living room
**Caption 3:** a large house shaped model is sitting in a living room
**Synthetic question 1:** what type of room has a tv in the center? **Answer:** living
**Synthetic question 2:** how large is the tv in the living room? **Answer:** small
**Question:** what era is this furniture from? **Predicted answer:** vintage

**Question:** what kind of sporting event is this? **GT answer:** soccer/not sure/pole vault
**Caption 1:** man on horse coming off from arena, holding something
**Caption 2:** a man is riding a horse during a soccer game
**Caption 3:** a man holding a red flag near a large person in a green field
**Synthetic question 1:** who is riding a horse in the middle of a stadium? **Answer:** man
**Synthetic question 2:** what color is the flag on display at a football game? **Answer:** red
**Question:** what kind of sporting event is this? **Predicted answer:** football

**Question:** what type of clouds are in the picture? **GT answer:** cumulus/cumuli/nimbus
**Caption 1:** a cloudy - filled sky on a cloudy day over a zebras
**Caption 2:** the clouds are gray and full of clouds
**Caption 3:** there are many different clouds in this sky
**Synthetic question 1:** what is in the background of a photo of a zebra? **Answer:** sky
**Synthetic question 2:** what type of sky is above on a cloudy day? **Answer:** cloudy
**Question:** what type of clouds are in the picture? **Predicted answer:** cloud

**Question:** how many people can this bus carry? **GT answer:** 50/40/39
**Caption 1:** a passenger bus traveling on a street side
**Caption 2:** blue commuter bus with parked on the side of the road
**Caption 3:** a bus that says aradara rides down the street
**Synthetic question 1:** what color bus is driving down the street? **Answer:** blue
**Synthetic question 2:** what is making it’s way down the street? **Answer:** bus
**Question:** how many people can this bus carry? **Predicted answer:** many

Figure 2. Failure case analysis for OK-VQA. Red color indicates incorrect prediction.
**Question**: which food has the least carbs? **GT answer**: soup/vegetable/salad

**Caption 1**: a table holding food including soup, sandwiches and fruit

**Caption 2**: the soup is very creamy in the bowl

**Caption 3**: sandwiches and soup is sitting on a table spread

**Synthetic question 1**: where is soup served on a table?

**Answer**: bowl

**Synthetic question 2**: what is on a plate next to a bowl of soup?

**Answer**: sandwich

**Question**: which food has the least carbs?

**Predicted answer**: soup

**Question**: in which way are the adults shown here likely related to the child? **GT answer**: parents/grandparents

**Caption 1**: a family sitting down on a bench in a park

**Caption 2**: a family sitting behind a park bench talking to a toddler

**Caption 3**: two people sitting on benches with a baby next to them

**Synthetic question 1**: what is sitting on a bench?

**Answer**: a baby

**Synthetic question 2**: who sits next to a toddler on a bench?

**Answer**: couple

**Question**: in which way are the adults shown here likely related to the child?

**Predicted answer**: parents

**Question**: what other surface is this game played on? **GT answer**: grass/clay/concrete

**Caption 1**: a blue surface with a blue tennis court

**Caption 2**: a man running across a blue tennis court with a racquet

**Caption 3**: a blue tennis court with a single game of tennis in progress

**Synthetic question 1**: what color is the tennis court?

**Answer**: blue

**Synthetic question 2**: what sport is a man playing on a blue court?

**Answer**: tennis

**Question**: what other surface is this game played on?

**Predicted answer**: grass

**Question**: what are they waiting to do when they stand next to the street? **GT answer**: cross/ride bus/light change

**Caption 1**: traffic and pedestrians at an intersection near a fire hydrant

**Caption 2**: a sidewalk and pedestrian crosswalk on a busy city street

**Caption 3**: a red fire hydrant stands besides a street that has a crosswalk

**Synthetic question 1**: where is a fire hydrant on a busy street?

**Answer**: crosswalk

**Synthetic question 2**: where are people waiting at a crosswalk?

**Answer**: intersection

**Question**: what are they waiting to do when they stand next to the street?

**Predicted answer**: cross

**Question**: what kind of resort are these people at? **GT answer**: ski resort/ski/snow

**Caption 1**: a group of people are skiing high up a slope

**Caption 2**: many people skiing down a ski slope during the day

**Caption 3**: a crowd of people on skis coming down the mountain

**Synthetic question 1**: what are people doing on a snow covered mountain?

**Answer**: ski

**Synthetic question 2**: who is skiing on a snow covered mountain?

**Answer**: people

**Question**: what kind of resort are these people at?

**Predicted answer**: ski resort

Figure 3. Success case analysis for A-OKVQA. Green color indicates answer cues and correct prediction.
Question: why is the woman holding the umbrella? GT answer: shade/sun protection/get shadow
Caption 1: a young woman and the umbrella are on an orange blanket
Caption 2: a woman's umbrella and two dogs under an umbrella
Caption 3: a woman holding an umbrella is getting some light under her umbrella
Synthetic question 1: who is holding an umbrella while her dog sits under it?
Answer: woman
Synthetic question 2: what is a woman holding and a dog under it?
Answer: an umbrella
Question: why is the woman holding the umbrella?
Predicted answer: to protect herself from the sun

Figure 4. Failure case analysis for A-OKVQA. Red color indicates incorrect prediction.
Figure 5. Success case analysis for VQAv2. Green color indicates answer cues and correct prediction.
Question: what is atop this building? GT answer: cross/stars/cross and stars
Caption 1: the cathedral tower is with the clock on a steeple
Caption 2: a clock and a two crosses on top of a church
Caption 3: the top of a red cathedral with a clock on the tower
Synthetic question 1: what part of a building has a clock on it?
Answer: top
Synthetic question 2: what color is the building with a clock on top?
Answer: red
Question: what is atop this building?
Predicted answer: a clock

Question: what are they standing by? GT answer: bushes/tree/bricks
Caption 1: two girl sitting and talking, one is looking at something
Caption 2: an older woman and young woman using cellphones
Caption 3: two girls sitting on a brick wall during the day time
Synthetic question 1: who are sitting on a bench looking at their phones?
Answer: women
Synthetic question 2: what type of wall are the two women sitting on?
Answer: brick
Question: what are they standing by?
Predicted answer: brick wall

Question: how many zebras are there? GT answer: 3
Caption 1: two zebras walking by a feeder full of food
Caption 2: pair of zebras standing together at water trough in zoo
Caption 3: the zebras are eating out of a feeder box
Synthetic question 1: how many zebras are standing next to each other?
Answer: two
Synthetic question 2: what are the zebras doing?
Answer: eating
Question: how many zebras are there?
Predicted answer: 2

Figure 6. Failure case analysis for VQAv2. Red color indicates incorrect prediction.
References


