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 $f_v^i \in \mathbb{R}^{K \times D_v^i}$ and question features as $f_q^i \in \mathbb{R}^{L \times D_q^i}$, 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_v^i is the dimension of patch feature in the i-th layer of ITE network and D_q^i 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

$$W^i = \text{softmax}\left(\frac{f_q^i W_Q^i {W_K^i}^\top {f_v^i}^\top}{\sqrt{D_q^i}}\right). \tag{1}$$

where $W_Q^i \in \mathbb{R}^{D_q^i \times D_q^i}$ is the query head and $W_K^i \in \mathbb{R}^{D_v^i \times D_q^i}$ 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 follwing 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_k^i, can be computed as the average over H attention heads and the sum over L textual tokens:

$$r_k^i = \frac{1}{H}\sum_{l=1}^{L}\sum_{h=1}^{H} \min\left(0, \frac{\partial\, \text{sim}(v, q)}{\partial W_{lk}^{ih}}\right) W_{lk}^{ih}, \tag{2}$$

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 OK-VQA 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.

Methods	A-OKVQA	
	Val	Test
Models Fine-Tuned in A-OKVQA Training Set		
Pythia [2]	25.2	21.9
ViLBERT [4]	30.6	25.9
LXMERT [7]	30.7	25.9
KRISP [5]	33.7	27.1
GPV-2 [3]	**48.6**	**40.7**
Zero-Shot Evaluation with Plug-in Frozen Large Language Model		
$\text{Ours}_{6.7B}$	33.3	32.2
Ours_{13B}	33.3	33.0
Ours_{30B}	36.9	36.0
Ours_{66B}	38.7	38.2
Ours_{175B}	42.9	**40.7**

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

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

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.

QA Pairs \ Caption	0	10	20	30	40	50
0	3.3	19.6	22.7	23.4	24.0	24.8
10	40.9	41.6	42.1	42.1	41.9	42.2
20	41.2	41.3	41.3	41.7	42.2	42.0
30	41.0	41.0	41.7	41.8	41.6	41.5
40	40.3	40.7	40.6	40.3	40.3	41.1
50	40.6	40.6	40.7	40.9	40.6	41.1

method is robust with different numbers of QA pairs and captions.

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.

Patch_num	10	20	40	Full
	41.2	41.8	41.6	39.8

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

Caption_num	PICa	10	30	50	100
	17.7	38.3	40.9	41.4	41.8

A.7. Examples

Question: what kind of bird are they? **GT answer:** seagull/pelican/seagul

Caption 1: two seagulls and a seagull on a wooden platform
Caption 2: a group of seagulls sit on some wood
Caption 3: a group of seagulls sitting down in the sunshine
Synthetic question 1: what birds are sitting on a wooden post?
Answer: seagulls
Synthetic question 2: how many seagulls are standing on top of a wooden post?
Answer: two
Question: what kind of bird are they?
Predicted answer: seagull

Question: what kind of beverage could one make with the item on top of the stove? **GT answer:** tea

Caption 1: a white kitchen with a stove, sink, and tea cups
Caption 2: kitchen with microwave, pots, coffee maker, stove and chairs
Caption 3: a kitchen filled with silver stove top oven sitting next to a microwave
Synthetic question 1: what is in the kitchen with a tea kettle?
Answer: stove
Synthetic question 2: what is on the counter next to the stove?
Answer: microwave
Question: what kind of beverage could one make with the item on top of the stove?
Predicted answer: tea

(b)

Question: what fabric are these jackets made of? **GT answer:** denim/jean

Caption 1: a man wearing a denims shirt stands at a motorcycle
Caption 2: man in denim jacket and blue uniform jacket on a red motorcycle
Caption 3: a man wearing blue denim clothes is standing near motorcycles
Synthetic question 1: what is a man wearing on a motorcycle?
Answer: a denim jacket
Synthetic question 2: what type of vehicle is the man sitting on?
Answer: motorcycle
Question: what fabric are these jackets made of?
Predicted answer: denim

(c)

Question: what style of fence is this? **GT answer:** picket/pickett

Caption 1: a fence of picket white boards with a gate
Caption 2: the house is fenced in in front of a white picketed fence
Caption 3: a white picket with pink roses in front of it
Synthetic question 1: what color is the picket fence in front of a house?
Answer: white
Synthetic question 2: what type of fence is in front of a house?
Answer: picket
Question: what style of fence is this?
Predicted answer: picket

(d)

Question: what is on the ears of the cattle in this photo? **GT answer:** tag

Caption 1: a row of cows, tied up to wires, yellow ears tags
Caption 2: a group of cows in grass with some yellow tags on their ears
Caption 3: cows with numbered ear tags standing behind a fence
Synthetic question 1: what are the cows wearing on their ears?
Answer: tags
Synthetic question 2: what color are the ear tags on the cows?
Answer: yellow
Question: what is on the ears of the cattle in this photo?
Predicted answer: tag

(e)

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 hot

Caption 1: two pizza boxes have pepper pizza and take out
Caption 2: two boxes are opened up of two different pizzas
Caption 3: there are two small baked pizzas on the table
Synthetic question 1: what are two large pizzas sitting in?
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

(b)

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

(c)

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

(d)

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

(e)

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

(b)

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

(c)

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

(d)

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

(e)

Figure 3. Success case analysis for A-OKVQA. Green color indicates answer cues and correct prediction.

Question: this dish is suitable for which group of people? **GT answer:** vegetarian/vegan/family

Caption 1: a pasta dish sitting on top of a white plate
Caption 2: a broccoli pasta dish that has very pasta
Caption 3: a dish of pasta with noodles and tomato sauce
Synthetic question 1: what vegetable is on a white plate?
Answer: broccoli
Synthetic question 2: what color is a plate of pasta with broccoli on it?
Answer: white
Question: this dish is suitable for which group of people?
Predicted answer: children

Question: what is in front of the monitor? **GT answer:** chair/keyboard/webcam

Caption 1: a corner table with computer computer on the desk
Caption 2: a computer on the small desk in a small office area
Caption 3: view of a computer monitor in a light lit room
Synthetic question 1: what is a computer sitting on in a corner of a room?
Answer: desk
Synthetic question 2: how big is the desk in the corner?
Answer: small
Question: what is in front of the monitor?
Predicted answer: desk

(b)

Question: what type of shot is the woman about to hit? **GT answer:** forehand/tennis shot/swing

Caption 1: tennis player is hitting a tennis ball with her racket
Caption 2: a woman in pink outfit hitting a tennis ball
Caption 3: a woman in a cropped top and pants swinging a tennis racquet
Synthetic question 1: what is a tennis player doing with a tennis racket?
Answer: swinging
Synthetic question 2: who is swinging a tennis racket at a tennis ball?
Answer: woman
Question: what type of shot is the woman about to hit?
Predicted answer: volley

(c)

Question: what is in the bottles? **GT answer:** alcohol/liqueur/baileys

Caption 1: a sandwich on a plate with a glass of beer bottle
Caption 2: a table that has a sandwich, beer, and beer on it
Caption 3: a sandwich on a plate with a glass of beer bottle
Synthetic question 1: what is next to a sandwich and a beer?
Answer: bottle
Synthetic question 2: where is a sandwich with a beer and beer on a plate?
Answer: table
Question: what is in the bottles?
Predicted answer: beer

(d)

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

(e)

Figure 4. Failure case analysis for A-OKVQA. Red color indicates incorrect prediction.

Question: what can the ram eat in this photo? **GT answer:** grass

Caption 1: the ram is standing outside on the green grass
Caption 2: a ram with white curly horns standing in a field
Caption 3: shaggy coated sheep with horns facing away in the center of a grass field
Synthetic question 1: where is a ram standing?
Answer: grass
Synthetic question 2: what animal is standing in a grassy field?
Answer: sheep
Question: what can the ram eat in this photo?
Predicted answer: grass

Question: what does the sign say? **GT answer:** stop

Caption 1: a stop sign with cloudy sky behind it
Caption 2: a red stop sign with a sky background
Caption 3: a tall stop sign on a rural road
Synthetic question 1: what color is the stop sign?
Answer: red
Synthetic question 2: what type of sky is behind a stop sign?
Answer: cloudy
Question: what does the sign say?
Predicted answer: stop

(b)

Question: what type animal is on the woman's pants? **GT answer:** owl/penguins

Caption 1: a girl is sitting on the ground in owl patterned pants
Caption 2: a woman with owly print pajamas pants is sitting in front of a pile of
Caption 3: a girl seated on the ground wearing pajamas
Synthetic question 1: where is a young girl wearing owl pants sitting?
Answer: the ground
Synthetic question 2: how is a young girl wearing owl pants doing?
Answer: sitting
Question: what type animal is on the woman's pants?
Predicted answer: owl

(c)

Question: how many children are at the table? **GT answer:** 3

Caption 1: three small little kids gather together on a dining table
Caption 2: a group of kids posing at a party table
Caption 3: three children sitting at a table with their food smiling at a picture
Synthetic question 1: what type of table are the three children sitting at?
Answer: dining
Synthetic question 2: how are the three children sitting at a table?
Answer: smiling
Question: how many children are at the table?
Predicted answer: 3

(d)

Question: is there broccoli in this dish? **GT answer:** yes

Caption 1: broccoli floret rice is in a large black pot
Caption 2: there is a closeup of a veggie salad
Caption 3: broccoli rice in a black bowl, ready to be eaten
Synthetic question 1: what is covered in broccoli in a pan?
Answer: rice
Synthetic question 2: what is a dish filled with broccoli and other vegetables in?
Answer: pot
Question: is there broccoli in this dish?
Predicted answer: yes

(e)

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

(b)

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

(c)

Question: how many buses are in the picture? **GT answer:** 8

Caption 1: a lot of buses sit parked in a line in front of a hill
Caption 2: a group of purple passenger buses all in a row
Caption 3: a row of purple bus buses next to each other
Synthetic question 1: how are the buses parked?
Answer: a line
Synthetic question 2: what color buses are parked in front of each other?
Answer: purple
Question: how many buses are in the picture?
Predicted answer: several

(d)

Question: are the numbers on the clock Roman numerals? **GT answer:** yes

Caption 1: a living room scene with a clock and tv
Caption 2: a chair is in front of a television that is being displayed
Caption 3: lounge chair with a clock that is hanging on the wall, and leather chair sits
Synthetic question 1: what is on in a living room?
Answer: television
Synthetic question 2: how is a wall clock displayed in a living room?
Answer: hanging
Question: are the numbers on the clock Roman numerals?
Predicted answer: no

(e)

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

References

[1] Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 930–945, Online, June 2021. Association for Computational Linguistics. 1

[2] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 2

[3] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. *arXiv preprint arXiv:2202.02317*, 2022. 2

[4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[5] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 2

[6] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*, 2022. 1

[7] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2