

# FROM IMAGES TO TEXTUAL PROMPTS: ZERO-SHOT VQA WITH FROZEN LARGE LANGUAGE MODELS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Large language models (LLMs) have demonstrated excellent zero-shot generalization to new tasks. However, effective utilization of LLMs for zero-shot visual question-answering (VQA) remains challenging, due to the modality disconnection and task disconnection between LLM and VQA task. End-to-end training on vision and language data may bridge the disconnections, but is inflexible and computationally expensive. To address this issue, we propose *Img2Prompt*, a plug-and-play module that provides the prompts that can bridge the aforementioned modality and task disconnections, so that LLMs can perform zero-shot VQA tasks without end-to-end training. In order to provide such prompts, we further employ LLM-agnostic models to provide prompts that can describe image content and self-constructed question-answer pairs, which can guide LLM to perform zero-shot VQA tasks. Img2Prompt offers the following benefits: 1) It is LLM-agnostic and can flexibly work with LLMs to perform VQA. 2) It renders end-to-end training unnecessary and significantly reduces the cost of deploying LLM for zero-shot VQA tasks. 3) It achieves comparable or better performance than methods relying on end-to-end training. On the challenging A-OKVQA dataset, our method outperforms some few-shot methods by as much as 20%.

## 1 INTRODUCTION

Visual question answering (VQA) (Antol et al., 2015) is a prominent vision-language task that finds a broad range of real-world applications, such as assisting blind individuals in understanding their environments. A diverse set of VQA datasets have been proposed, some focusing on image recognition (Goyal et al., 2017; Antol et al., 2015) and others on logical reasoning (Marino et al., 2019). However, the human annotation is expensive to obtain and may introduce a variety of human biases (Changpinyo et al., 2022; Banerjee et al., 2021; Yuan, 2021), making the VQA system brittle towards new answer styles and question types (Agrawal et al., 2018; Kafle & Kanan, 2017). This has led researchers to zero-shot VQA methods (Changpinyo et al., 2022; Banerjee et al., 2021; Kafle & Kanan, 2017) that do not require ground-truth question-answer annotations, thereby facilitating more generalizable VQA systems.

Recently, large language models (LLMs) (e.g., (Brown et al., 2020a; Zhang et al., 2022)) have demonstrated excellent capabilities to perform tasks with zero in-domain data, conduct logical reasoning, and apply commonsense knowledge in NLP tasks (Kojima et al., 2022; Wei et al., 2022b;c). As a result, recent approaches (Alayrac et al., 2022; Yang et al., 2022; Tsimpoukelli et al., 2021) have resorted to leverage LLMs in zero-shot VQA.

However, applying LLMs to VQA tasks is less than straightforward. Degraded performance results from the modality disconnect between vision and language, and the task disconnect between language modeling and question answering. A commonly used technique is to finetune a vision encoder jointly with the LLM (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Jin et al., 2022), which aligns the vision and language representation spaces. The finetuning is computationally expensive. For example, Flamingo (Alayrac et al., 2022) trains layers newly inserted into the LLM on billions of image-text pairs with thousands of TPUs. Further, the training specializes the vision encoder and the LLM and introduces strong interdependence between them. Thus, if we want to replace the two networks as improved pretrained models emerge, we must perform the finetuning again.

In contrast to the computational costly end-to-end integration of LLM into a VQA system, we expect that, from the perspective of general-purpose AI and industrial deployment, the LLM is able to perform VQA task off-the-shelf, which brings two benefits. First, it can reduce the deployment cost and simplifies the deployment. Second, in case we want to update the LLM, doing so is straightforward. Nevertheless, a major concern is that using LLM off-the-shelf in zero-shot VQA without end-to-end training may not achieve good performance, as the system may not cope with the modality disconnect and the task disconnect properly. Yang et al. (2022) bridge the task disconnect by providing exemplar QA pairs from training data as prompt to the LLM, but doing so assumes the existence of training data and is sensitive to the selection of few-shot samples

We propose *Img2Prompt*, a plug-and-play module that enables LLM to perform zero-shot VQA off-the-shelf. This eliminates the need for the expensive end-to-end vision-language representation alignment, thereby allowing low-cost and flexible model deployment. The central insight of Img2Prompt is that we can utilize the vision-language model, e.g. BLIP (Li et al., 2022), and text question generation model to pretranslate the image content into exemplar question-answer (QA) pairs, which are fed to the LLM as part of the prompt. These exemplars tackle the modality disconnect by describing the image content verbally, and tackle the task disconnect by demonstrating the QA task to the LLM. We apply Img2Prompt to the open-sourced OPT language model Zhang et al. (2022) to perform zero-shot VQA task. Experiments show that Img2Prompt enables OPT in different sizes to achieve comparable or even superior zero-shot VQA performance to methods that perform costly end-to-end training.

**Contributions:** a) We propose Img2Prompt, a plug-and-play module that generates synthetic question-answer pairs from the current image of the question, which is able to bridge the modality disconnection and the task disconnection between language modeling and visual question-answering. b) We demonstrate empirically that a large language model is able to be used off-the-shelf for zero-shot VQA tasks without costly end-to-end training or specialized textual QA networks. Doing so reduces model deployment costs and offers the flexibility of model updates. c) Our experimental results show that the OPT models equipped with Img2Prompt achieve zero-shot VQA performance that is competitive or superior to the end-to-end trained models. For example, we outperform Flamingo Alayrac et al. (2022) by 5.6% on VQAv2. We even outperform some few-shot VQA methods.

## 2 RELATED WORK

### 2.1 RECENT ADVANCES IN VQA METHODS

As a multi-modal evaluation benchmark, Visual Question Answering (VQA) that requires the model to answer a natural language question according to the image, had been actively studied (Yang et al., 2016; Anderson et al., 2018; Antol et al., 2015; Schwenk et al., 2022; Akula et al., 2021). With vision-language models (Jiang et al., 2022; Yuan et al., 2021; Lu et al., 2019; Li et al., 2021; 2020b; Zhang et al., 2021; Wang et al., 2022; Singh et al., 2021; Li et al., 2022; Jin et al., 2022; Dai et al., 2022) that are pretrained on large-scale image-text datasets, VQA tasks have advanced rapidly through the fine-tuning of these models in specific VQA dataset. In order to solve knowledge-based VQA (Schwenk et al., 2022; Marino et al., 2019), recent works (Gui et al., 2021; Lin et al., 2022; Wu et al., 2022; Luo et al., 2022; 2021; Marino et al., 2021; Gardères et al., 2020; Li et al., 2020a) incorporate external knowledge, such as ConceptNet (Speer et al., 2017) or Wikipedia, into the networks, but experimental results in (Schwenk et al., 2022) show that these methods still struggle to answer questions requiring reasoning ability.

### 2.2 LLM FOR ZERO/FEW-SHOT VQA TASKS

Large language models (LLMs) (Brown et al., 2020b; Zhang et al., 2022; Chowdhery et al., 2022) trained on web-scale corpus are powerful in natural language understanding and reasoning (Zhou et al., 2022; Brown et al., 2020a). To infer on task data, LLMs typically generate target tokens autoregressively. In specific, given prompt  $C$  and task input  $x$ , an LLM generates target tokens  $Y = \{y_i\}_{i=1}^n$ , with  $y_i = \arg \max p_\theta(y_i | y_{<i}, C, x)$  and  $\theta$  the model parameters. Prior VQA methods using LLMs exist, and mainly fall into two categories: multi-modal pretraining and language-mediated VQA, as reviewed below.

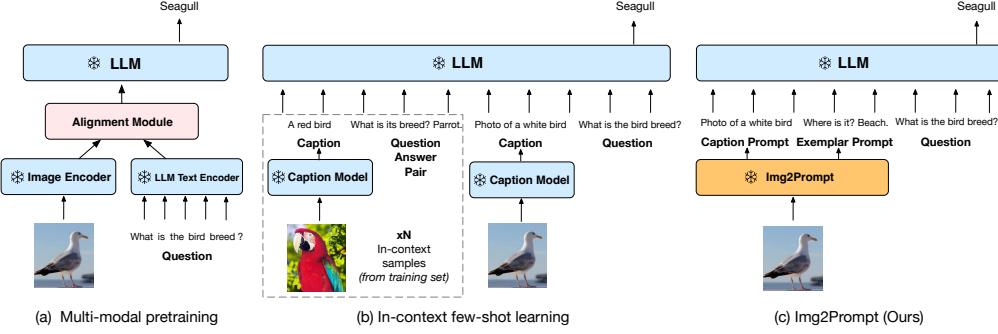


Figure 1: The illustrative comparison of three types of methods that enable LLM to perform VQA tasks, where blue block denotes that the inner parameters are frozen while pink block indicates the inner parameters are trainable.

**Multi-modal pretraining.** These approaches align vision and language embeddings by training additional alignment modules, as shown in Figure 1(a). Considering that LLMs are too large to fine-tune efficiently, (Tsimpoukelli et al., 2021) opt to fine-tune only the visual encoder while Flamingo (Alayrac et al., 2022) trains extra cross-attention layers to model cross-modality interactions. However, this paradigm suffers from two drawbacks: 1) Highly compute-inefficient. Jointly aligning vision backbones and LLMs requires large compute resources. For example, training Flamingo requires 1536 TPUv4 over two weeks. Hence, it becomes prohibitively expensive to switch to a different LLM. 2) Catastrophic forgetting. The alignment step may be detrimental to LLMs’ reasoning ability, if the LLMs are jointly trained with the visual model (Alayrac et al., 2022).

**Language-mediated VQA.** Instead of vectorized representations, this VQA paradigm directly resorts to natural language as the intermediate representation of the image and no longer requires expensive pretraining. As depicted by Figure 1(b), it first converts the current image to language descriptions and feeds the descriptions, possibly accompanied by in-context exemplars, to a frozen LLM. In a few-shot setting, PICa (Yang et al., 2022) generates captions for the image and selects training data samples as in-context exemplars, but its performance degrades substantially when the exemplars are omitted. As a concurrent zero-shot approach, (Anonymous, 2022) generates question-relevant captions. Due to the zero-shot requirement, it is unable to provide in-context exemplars and does not reap the benefits of in-context learning. As a result, it has to rely on a QA-specific LLM, UnifiedQAv2 (Khashabi et al., 2020), to achieve high performance.

### 3 METHOD

Difficulties in utilizing LLMs effectively in zero-shot VQA stem mainly from two obstacles: (i) *The modality disconnection*: LLMs do not natively process images and encoding visual information into a format that LLMs can process can be a challenge. (ii) *The task disconnection*: LLMs are usually pretrained using generative (Brown et al., 2020a) or denoising objectives (Devlin et al., 2019) on language modeling tasks. As the LLMs are unaware of the tasks of question answering or VQA, they often fail to fully utilize contextual information in generating the answers.

In language-mediated VQA (Yang et al., 2022; Meng Huat Tieng et al., 2022), the modality disconnection is addressed by converting the image to intermediate language descriptions instead of dense vectors (§2.2). The task disconnection must be addressed using either few-shot in-context exemplars (Yang et al., 2022) or a directly finetuned on textual QA model (Meng Huat Tieng et al., 2022). It is not clear how to tackle the task disconnection on generic LLMs under zero-shot settings.

We propose a new zero-shot technique to address the task disconnection on generic LLMs, Img2Prompt (Figure 1c), which generates image-relevant exemplar prompts for the LLM. Our key insight is that we can generate synthetic question-answer pairs as in-context exemplars from the *current* image of the question. The exemplars not only demonstrate the QA task but also communicate the content of the image to the LLM, thereby hitting two birds with one stone. Img2Prompt is LLM-agnostic; it unlocks the knowledge and the reasoning capacity of off-the-shelf LLMs, offering a powerful yet flexible solution for zero-shot VQA.

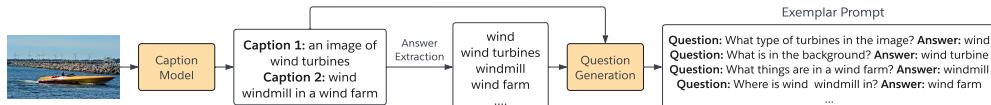


Figure 2: The illustration of answer extraction and question generation.

Table 1: Results from mixing captions and exemplar prompts on 30B OPT (Zhang et al., 2022).

Prompt Template	Caption Prompt	Exemplar Prompt	VQAv2 val	OK-VQA
Instruction	✗	✗	18.1	3.3
Instruction + Captions	✓	✗	46.1	23.5
Instruction + Question-Answer Pairs	✗	✓	57.9	41.1
Instruction + Captions + Question-Answer Pairs	✓	✓	59.5	41.8

### 3.1 ANSWER EXTRACTION

In order to incorporate the image content into the exemplars for in-context learning, from the current VQA image, we first seek words that could serve as answers to synthetic questions. We generate a number of captions using an off-the-shelf question-relevant caption generation module (§3.3). Following recent papers (Changpinyo et al., 2022; Lee et al., 2021), we extract noun phrases (including named entities), verb phrases, adjective phrases, numbers, and boolean-typed words like “yes” and “no” as potential answers<sup>1</sup>. We show some extracted answer candidates in Figure 2 and Figure A.7 in the Appendix.

### 3.2 QUESTION GENERATION

With the extracted answer candidate set  $\{\hat{a}_j\}_{j=1}^U$ , we can directly use any question generation network (Kafle et al., 2017; Lu et al., 2020; Xu et al., 2020; Kil et al., 2021; Akula et al., 2021) to generate specific questions for each answer candidate. In this paper, we try both template-based generation and neural networks. Note that to avoid violating the zero-shot requirement, we opt not to use question generation models specifically designed for VQA (Li et al., 2018; Vedd et al., 2021).

**Template-based Question Generation.** Using an off-the-shelf parser, we differentiate the part of speech of each answer, and design specific question templates for each type. For example, for answers that are nouns, we use the question “What object is in this image?” For verb answers, we use the question “What action is being taken in this image?” Due to space constraints, we leave the complete list of templates to Appendix A.5.

**Neural Question Generation.** Inspired by (Changpinyo et al., 2022), we train a neural question generation model on textual QA datasets. Specifically, we finetune a pretrained T5-large model (Raffel et al., 2020) to generate questions from answers. The input to the model contains the prompt “Answer: [answer]. Context: [context]”, where [answer] denotes the answer text and [context] denotes the context text from textual QA datasets. During inference, we replace [answer] with an extracted answer candidate and [context] with the generated caption from which the answer was extracted. The model is finetuned on five textual QA datasets including SQuAD2.0 (Rajpurkar et al., 2018), MultiRC (Khashabi et al., 2018), BookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2018) and Social IQA(Sap et al., 2019).

With the above question generation methods, we acquire a set of synthetic question-answer pairs  $\{\hat{q}_j, \hat{a}_j\}_{j=1}^U$ . We use these question-answer pairs as exemplars of LLM in-context learning (Brown et al., 2020a), which guides the LLM to perform QA task given the image content and bridges the task disconnect between language modeling and VQA.

As a sneak preview, we show effects of exemplar QA pairs in Table 1. The details of the instructions are explained in §4.5. We observe that exemplar QA prompts perform considerably better than caption prompts (detailed in §3.3) only, demonstrating their efficacy in bridging the task disconnection between LLM pre-training and VQA tasks. Moreover, since the exemplar prompts already describe much content of the image, which helps to bridge the modality disconnection, adding captions on top does not provide much new information and brings only limited performance gains.

<sup>1</sup>We use the spaCy parser at <https://spacy.io/>, though are not tied to the parser in any way.

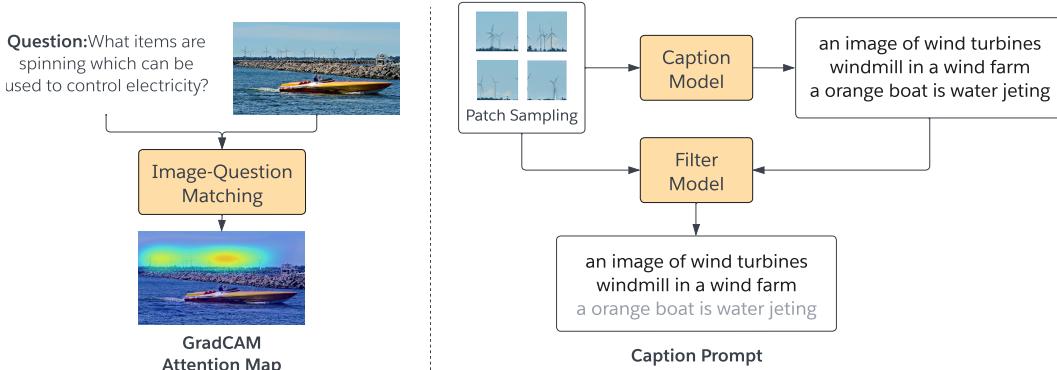


Figure 3: The generation process of Caption Prompt in Img2Prompt module, where Image-Question Matching Model and Caption Filter Model are Image-grounded Text Encoder in BLIP, and Caption Model is also from BLIP. The left part refers to (Meng Huat Tiong et al., 2022).

### 3.3 QUESTION-RELEVANT CAPTION PROMPT

In addition to the synthetic exemplar QA pairs, we also supply question-relevant image captions to the LLM. We observe that the question may ask about specific objects or regions in the image (Wu et al., 2019) but generic captions generated by existing networks may not contain relevant information. In Figure 3, the question “*What items are spinning in the background which can be used to control electricity?*” is relevant only to the wind turbines. However, captions generated from the complete image are likely to focus on the salient orange boat, leaving LLM with no information to answer the question. To address this issue, we generate captions about the question-relevant portion of the image and include them in the prompt to the LLM.

To achieve this, we first determine the regions of the image that are relevant to the question, by using the Image-grounded Text Encoder (ITE) in BLIP (Li et al., 2022), as which assigns a similarity score  $\text{sim}(v, q)$  to any pair of image  $v$  and textual question  $q$ . With ITE, we use GradCAM (Selvaraju et al., 2017), a feature-attribution interpretability technique, to generate a coarse localisation map highlighting matching image regions given a question (Li et al., 2022). Briefly, GradCam qualifies the cross-attention scores from the Transformer network by the gradient of ITE similarity function  $\text{sim}(v, q)$  with respect to the cross-attention scores. As this technique was proposed in (Meng Huat Tiong et al., 2022), we leave the details to Appendix A.2.

Having obtained the patch relevance  $r$ , we sample a subset of image patches with probability proportional to patch relevance  $r$ . After that, we generate captions from the sampled image patches using top-k sampling (Fan et al., 2018). To generate semantically meaningful captions, a short prompt, “a picture of,” is also fed into the text decoder. We repeat this process  $M$  times for each image to generate  $M$  diverse captions, and keep only captions that are not exact substrings of others.

However, due to the non-deterministic nature of top-k sampling, the caption model may generate noisy captions that have a negative impact on performance. To remove noisy captions, we use ITE to calculate the similarity score between the generated caption and sampled question-relevant image patches, and filter captions with less than 0.5 matching scores. Overall, this process yields synthetic captions that are question-relevant, diverse, and clean, providing a bridge between visual and language information.

### 3.4 PROMPT DESIGN

With synthetic question-relevant captions and question-answer pairs, we construct complete prompts for LLM by concatenating the instruction, captions, and QA exemplars. The instruction text is “Please reason the answers of question according to the contexts.” The caption prompt is formatted as “Contexts: [all captions]”. Individual QA exemplars are formatted as “Question: [question] Answer: [answer]” and concatenated. We position the current question as the last portion of the prompt, formatted as “Question: [question]. Answer: ”. Finally, to get the answer, we perform greedy decoding on the LLM and remove meaningless tokens as in Flamingo.

Furthermore, as the input to LLMs has maximum lengths, e.g. 2048 in OPT and GPT3, it is necessary to select a subset of question-relevant captions and question-answer pairs to construct the

prompt. To select the most informative prompt, we first count the frequency of the synthetic answer candidates in 100 generated captions. We then select 30 answer candidates with highest frequencies and generate one question for each. Also, we include 30 answers with the lowest frequency and one caption containing each answer. See §4.4 for analysis of caption selection strategies.

## 4 EXPERIMENT

In this section, we first validate the efficacy of Img2Prompt by comparing it with other zero-shot and few-shot VQA methods. Then, we perform ablation studies on important design choices, such as prompt patterns and caption selection strategies, to understand their effect. We also show qualitative examples and include discussion on observed failure cases.

### 4.1 ENVIRONMENT SETUP

**Datasets.** We validate our method on VQAv2 (Goyal et al., 2017), OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) datasets, which contain questions requiring perception, reasoning and commonsense to answer. Specifically, VQAv2 (Goyal et al., 2017) contains 214,354 questions in the validation set and 107,394 in the test-dev dataset. OK-VQA (Marino et al., 2019) and A-OK-VQA (Schwenk et al., 2022) emphasize on commonsense reasoning, among which OK-VQA contains 5,046 test questions and A-OKVQA (Schwenk et al., 2022) contains 1,100 validation questions and 6,700 test questions.

**Implementation details.** To obtain question-relevant caption prompt (§3.3), we use BLIP (Li et al., 2022) to generate captions and perform image-question matching. To localize the image regions relevant to the question, we generate GradCam from the cross-attention layer of BLIP image-grounded text encoder. We then sample  $K' = 20$  image patches based on GradCam, and use them to obtain 100 question-relevant captions. We use OPT as LLMs, and experiment with its 6.7B, 13B, 30B, 66B, 175B variants. In this way, we show that our method generalizes to LLMs of different scales. **We also report the results of different LLMs in Appendix A.3.** We use LLMs to generate answers autoregressively, without access to either answer list or training samples, thereby facilitating zero-shot VQA. We follow official evaluation protocols and report VQA scores on each dataset.

**Competing methods.** We compare with prior VQA methods, which roughly fall into three categories: (i) *Zero-shot methods with frozen LLMs*, such as PICa (Yang et al., 2022). Our method also belongs to this category, yet unlike PICa, Img2Prompt requires no training samples to compose the prompts. (ii) *Zero-shot methods with extra multi-modal pre-training*, such as Flamingo (Alayrac et al., 2022), Frozen (Tsimpoukelli et al., 2021), VL-T5 (Cho et al., 2021), FewVLM (Jin et al., 2022) and VLKD (Dai et al., 2022). These methods require large-scale vision-language datasets and are costly to update. We also include results from VQ<sup>2</sup>A (Changpinyo et al., 2022) and WeaQA (Banerjee et al., 2021) in this category, with *caveats* that they assume access to answer candidates which may not be available in practice. Therefore, their results should be interpreted with caution. (iii) For reference purposes, we also include available results from *few-shot methods*. These include few-shot results of PICa (Yang et al., 2022), FewVLM (Jin et al., 2022) and ClipCap (Mokady et al., 2021).

### 4.2 MAIN RESULTS

Main quantitative results are shown in Table 2. We summarize our findings as follows.

**State-of-the-art results on zero-shot evaluation with plug-in LLMs.** Img2Prompt surpasses PICa, the best prior zero-shot model with frozen LLMs, by a significant margin (17.7 *versus* 45.6 on OK-VQA), thereby establishing a new state-of-the-art. In addition, we remark that despite PICa uses frozen LLMs, it requires training samples to build prompts. In contrast, our method generates question-answers with no access to VQA samples, thus fully fulfilling the zero-shot requirements.

**Scaling effect of LLMs and their emergent capabilities on VQA.** When increasing the number of parameters of LLMs from 6.7B to 175B, we see a 3-10 points improvement in VQA scores across datasets. This shows that stronger language modeling capabilities help better comprehend the question, thus giving more accurate answers. Such a trend is even more clear and consistent on OK-VQA and A-OKVQA, whose questions demand commonsense reasoning and external knowledge that LLMs excel at providing. This corroborates our belief that LLMs are beneficial to VQA.

Table 2: Performance on VQAv2, OK-VQA, and A-OKVQA. A few methods do not strictly satisfy the zero/few-shot requirements: methods without end-to-end training but assumes access to training samples are labeled with  $\dagger$ ; methods that answer from a predefined list of candidates are in grey. Further,  $\times$  annotates methods requiring no end-to-end training, which is desirable, and  $\checkmark$  otherwise.

Methods	End-to-End Training?	Shot Number	VQAv2 val	VQAv2 test	OK-VQA val	A-OKVQA val	A-OKVQA test
<i>Zero-Shot Evaluation with Frozen Large Language Model</i>							
PICa <sub>175B</sub> $\dagger$	$\times$	0	-	-	17.7	-	-
Img2Prompt <sub>6.7B</sub>	$\times$	0	57.6	57.0	38.2	33.3	32.2
Img2Prompt <sub>13B</sub>	$\times$	0	57.1	57.3	39.9	33.3	33.0
Img2Prompt <sub>30B</sub>	$\times$	0	59.5	60.4	41.8	36.9	36.0
Img2Prompt <sub>66B</sub>	$\times$	0	59.9	60.3	43.2	38.7	38.2
Img2Prompt <sub>175B</sub>	$\times$	0	<b>60.6</b>	<b>61.9</b>	45.6	<b>42.9</b>	<b>40.7</b>
<i>Zero-Shot Evaluation with Extra End-to-End Training</i>							
VL-T5 <sub>no-vqa</sub>	$\checkmark$	0	13.5	-	5.8	-	-
FewVLM <sub>base</sub>	$\checkmark$	0	43.4	-	11.6	-	-
FewVLM <sub>large</sub>	$\checkmark$	0	47.7	-	16.5	-	-
VLKD viT-B/16	$\checkmark$	0	38.6	39.7	10.5	-	-
VLKD viT-L/14	$\checkmark$	0	42.6	44.5	13.3	-	-
Frozen <sub>7B</sub>	$\checkmark$	0	29.5	-	5.9	-	-
Flamingo <sub>3B</sub>	$\checkmark$	0	-	49.2	41.2	-	-
Flamingo <sub>9B</sub>	$\checkmark$	0	-	51.8	44.7	-	-
Flamingo <sub>80B</sub>	$\checkmark$	0	-	56.3	<b>50.6</b>	-	-
<i>Zero-shot Evaluation with Access to Answer Candidates</i>							
WeQA ZSL	$\checkmark$	0	46.8	-	-	-	-
VQ <sup>2</sup> A	$\checkmark$	0	61.1	-	19.8	-	-
<i>Few-Shot Evaluation</i>							
ClipCap $\rightarrow$ Cap $\rightarrow$ GPT <sub>175B</sub>	$\times$	10	-	-	-	16.6	15.8
ClipCap $\rightarrow$ Rel $\rightarrow$ GPT <sub>175B</sub>	$\times$	10	-	-	-	18.1	15.8
FewVLM <sub>base</sub>	$\checkmark$	16	48.2	-	15.0	-	-
FewVLM <sub>large</sub>	$\checkmark$	16	51.1	-	23.1	-	-
PICa <sub>175B</sub> $\dagger$	$\times$	1	-	-	36.4	-	-
PICa <sub>175B</sub> $\dagger$	$\times$	4	-	-	43.3	-	-
PICa <sub>175B</sub> $\dagger$	$\times$	16	54.3	-	46.5	-	-
PICa <sub>175B</sub> -Ensemble	$\times$	80	56.1	-	48.0	-	-

Another intriguing phenomenon we observe is that the effect of scaling LLMs becomes obvious only when the model size becomes sufficiently large, for example, when using 30B or larger models, while not entirely predictable on smaller ones (6.7B and 13B). This echoes with the recent finding on the emergent abilities when using LLMs off-the-shelf (Wei et al., 2022a) for language tasks, while confirming the same trend for the first time when using frozen LLMs for vision(-language) tasks.

**Competitive performance with end-to-end pretraining and few-shot models.** Img2Prompt obtains superior performance to most models with end-to-end pretraining, as well as those evaluated in few-shot setups. For example, on VQAv2 our method surpasses Flamingo<sub>80B</sub>, which cost over 500K TPU hours and billion-scale datasets to train, by a margin of 5.6 points. On A-OKVQA, Img2Prompt more than doubles the best reported results so far, from ClipClap. The only a few exceptions are on OK-VQA, where our method obtains better results than Flamingo<sub>9B</sub>, yet is not able to stay on par with Flamingo<sub>80B</sub>. Considering that Img2Prompt is flexible to adapt to updated and stronger LLMs with zero extra training cost, we consider it a more approachable solution to practical adoption of VQA systems, than those trained end-to-end. We also include comparisons with supervised models in Appendix A.4. Img2Prompt achieves better performance than most supervised models, despite the fact that it uses zero training data and evaluated in a zero-shot setup. These results once again validates its effectiveness.

#### 4.3 ANALYSIS ON QUESTION GENERATION METHODS

Table 3 shows the performance of different question selection strategies described in Section 4.5. We compare three question generation techniques, include *image-agnostic*, which uses questions sam-

pled from other images; *template*-based, which uses template questions, and *neural*-based, which uses neural generated questions. Further, we compare two synthetic QA selection strategies. The *random* strategy, which selects QA pairs for prompt randomly; the *max freq.* approach, which selects answer candidates that are most frequent in the captions, and also retrieve the associated synthetic questions to build the prompt.

Among the three question generation techniques, *Agnostic* perform the worst whereas *Neural* performs the best. We attribute the differences to the quality of QA pairs. *Agnostic* QA pairs contain information irrelevant to the current image and may mislead the LLM. *Template* questions feature little linguistic variation and hence cannot demonstrate different QA strategies. *Neural* has the most relevant information and the most linguistic diversity. QA pair with maximum answer frequency outperform random questions. We hypothesize that the most frequent answers describe the most salient aspects of the image, thereby providing more information than random questions.

In addition, we evaluate visual information quality encoded in the exemplar prompts using the answer hit rate and the answer noise rate. Answer hit rate (AHR) is defined as the proportion of QA pairs containing the ground-truth answer. Answer noise rate (ANR) is defined as the ratio of ground-truth answers to the total number tokens in the exemplar prompts. Table 4 indicates that exemplar prompts generated from question-relevant captions have a higher AHR, hence enhancing the VQA performance. In addition, the caption filter procedure can remove some noisy captions, allowing it to achieve a higher ANR than its competitors. The experimental results demonstrate that improving both the AHR and the ANR can improve the quality of prompts and VQA performance.

#### 4.4 ABLATION ON CAPTION SELECTION

As Table 6 shows, we evaluate the performance different caption selection strategies, where Max Frequency selects captions containing 30 answers with highest frequencies and Min Frequency selects answers with the lowest frequencies. As the exemplar prompts are produced with answers with the highest frequencies, the Max Frequency strategy does not provide more information than exemplar prompts. In contrast, the Min Frequency strategy chooses captions that can provide some information not in the QA pairs, providing a performance boost.

#### 4.5 ABLATION STUDY ON PROMPT DESIGN

We have two options to construct LLM’s prompt. The first option is to append a synthetic QA pair after the caption that the QA pair is generated from. This can be described as CQA-CQA-CQA, where C, Q, A stand for caption, synthetic question, and synthetic answer respectively. Alternatively, we can present all captions at once, followed by all question-answer pairs, which we denote as CCC-QAQQA. Experimentally (Table 5), the second design performs significantly better than the first. We hypothesize that the first design may induce the LLM to read only one caption before answering, since in the prompt this caption contains all the information needed for the question. While it is hard to pinpoint the actual mechanism, the results highlight the importance of QA prompts and their positions.

Table 4: The experimental results on QA pairs generated from different captions. The results are run with OPT 30B.

Exemplar Prompts Generation Source	OK-VQA			VQAv2 val		
	VQA Score	Answer Noise Rate	Answer Hit Rate	VQA Score	Answer Noise Rate	Answer Hit Rate
Caption from Complete Image	39.8	0.018	0.480	57.1	0.0290	0.725
Question-relevant Caption	40.6	0.022	<b>0.581</b>	58.1	0.0303	<b>0.821</b>
Question-relevant Caption with Filter	<b>41.8</b>	<b>0.025</b>	0.566	<b>59.5</b>	<b>0.0313</b>	0.804

Table 5: Ablations on prompts designs.

Methods	OK-VQA	VQAv2 val
CQA-CQA-CQA	37.8	52.1
CCC-QAQQA	41.8	59.5

Table 6: Ablation on caption selection methods.

Caption Selection	Random	Max Frequency	Min Frequency
	41.3	41.1	<b>41.8</b>

#### 4.6 EXAMPLES AND FAILURE CASE ANALYSIS

In Figure 4, we show four examples of caption and exemplar prompts and the predictions, including cases of success and failure. In Figure 4(a), the captions and the synthetic QA pairs provide the information that a man is making drinks at a bar. The LLM draws on background knowledge and correctly infers that his job is bartender. In Figure 4(c), while the prediction is understandable (even if not strictly grammatical), the LLM is unable to make inferences based on qualitative physics and predict the right answer. These results highlight the importance to apply appropriate commonsense knowledge in open-ended VQA.

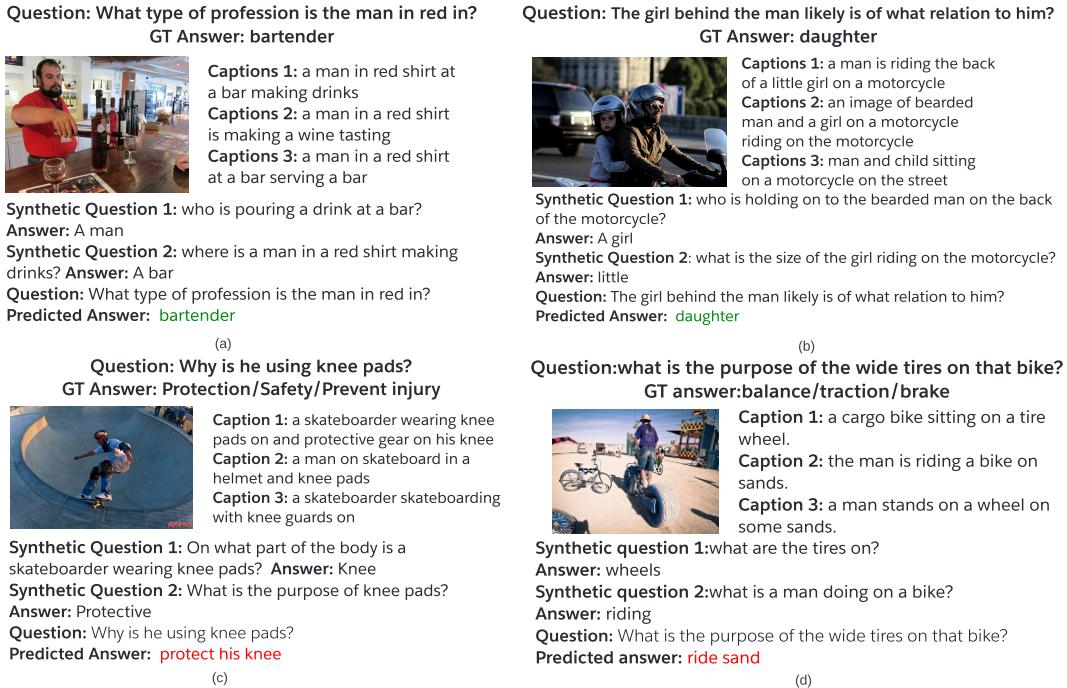


Figure 4: Example predictions made by Img2Prompt. Specifically, (a) and (b) are successful cases, while (c) and (d) are failure cases. See more examples at Appendix A.7.

#### 5 LIMITATION

One limitation of the proposed approach is that generating image captions and question-answer pairs incurs extra inference overhead. On an  $8 \times A100$  machine, our current implementation brings about 24.4% additional computational time on top of the inference time of 175B OPT. We note that further reduction of the overhead can be obtained by shortening the prompt, trading accuracy for speed. Additionally, our method avoids expensive end-to-end multimodal representation alignment, which, in the case of Flamingo, took more than 500K TPU hours.

#### 6 CONCLUSION

In this paper, we propose Img2Prompt, a plug-and-play module designed to exploit the knowledge and reasoning power of large language models (LLMs) off-the-shelf for zero-shot VQA tasks. Concretely, Img2Prompt provides visual information and task guidance to LLMs in the format of easily-digestible prompts. This eliminates the requirement for the expensive end-to-end vision-language alignment, increasing model deployment flexibility while decreasing model deployment cost. The experiments show that Img2Prompt enables different LLMs to achieve comparable or even superior zero-shot VQA performance to other methods that require costly end-to-end training.

## 7 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. We will publicly release all codes after the acceptance of this paper.

## REFERENCES

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.
- Arjun Akula, Soravit Changpinyo, Boqing Gong, Piyush Sharma, Song-Chun Zhu, and Radu Soricut. Crossvqa: Scalably generating benchmarks for systematically testing vqa generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2148–2166, 2021.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *arXiv Preprint 2204.14198*, 2022. URL <https://arxiv.org/abs/2204.14198>.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.
- Anonymous. Plug-and-play VQA: Zero-shot vqa by conjoining foundation models with zero training. *Anonymous*, 2022.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. WeaQA: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3420–3435, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.302. URL <https://aclanthology.org/2021.findings-acl.302>.
- 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*, pp. 930–945, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.72. URL <https://aclanthology.org/2021.naacl-main.72>.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL <https://doi.org/10.5281/zenodo.5297715>. If you use this software, please cite it using these metadata.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot

- learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020b.
- Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *North American Chapter of the Association for Computational Linguistics*, 2022. doi: 10.48550/ARXIV.2205.01883. URL <https://arxiv.org/abs/2205.01883>.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1931–1942. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/cho21a.html>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2383–2395, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.187. URL <https://aclanthology.org/2022.findings-acl.187>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL <https://aclanthology.org/P18-1082>.
- François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 489–498, 2020.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6904–6913, 2017. URL [https://openaccess.thecvf.com/content\\_cvpr\\_2017/html/Goyal\\_Making\\_the\\_v\\_CVPR\\_2017\\_paper.html](https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html).
- Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021.
- Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15513–15523, 2022.
- 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.

- Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2763–2775, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.197. URL <https://aclanthology.org/2022.acl-long.197>.
- Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pp. 1965–1973, 2017.
- Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 198–202, 2017.
- 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.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 252–262, 2018.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020. URL <https://arxiv.org/abs/2202.12359>.
- Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. *arXiv preprint arXiv:2109.06122*, 2021.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.
- Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. Qace: Asking questions to evaluate an image caption. *arXiv preprint arXiv:2108.12560*, 2021.
- Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1227–1235, 2020a.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf>.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. URL <https://arxiv.org/abs/2201.12086>.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pp. 121–137. Springer, 2020b. doi: 10.1007/978-3-030-58577-8\8. URL [https://doi.org/10.1007/978-3-030-58577-8\\_8](https://doi.org/10.1007/978-3-030-58577-8_8).
- Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6116–6124, 2018.

- Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. Revive: Regional visual representation matters in knowledge-based visual question answering. *arXiv preprint arXiv:2206.01201*, 2022.
- 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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf>.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10437–10446, 2020.
- Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *arXiv preprint arXiv:2109.04014*, 2021.
- Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. Vc-gpt: Visual conditioned gpt for end-to-end generative vision-and-language pre-training. *arXiv preprint arXiv:2201.12723*, 2022.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.
- 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*, pp. 14111–14121, 2021.
- Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv e-prints*, pp. arXiv–2210, 2022.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.
- Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.
- Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*, 2022.
- 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.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.
- Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.
- 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)*, pp. 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL <https://aclanthology.org/D19-1514>.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 200–212. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/file/01b7575c38dac42f3cfb7d500438b875-Paper.pdf>.
- Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. Guiding visual question generation. *arXiv preprint arXiv:2110.08226*, 2021.
- Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022. URL [https://openreview.net/forum?id=GUrhfTuf\\_3](https://openreview.net/forum?id=GUrhfTuf_3).
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022b.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022c.
- Jialin Wu, Zeyuan Hu, and Raymond J Mooney. Generating question relevant captions to aid visual question answering. *arXiv preprint arXiv:1906.00513*, 2019.
- Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2712–2721, 2022.
- Xing Xu, Tan Wang, Yang Yang, Alan Hanjalic, and Heng Tao Shen. Radial graph convolutional network for visual question generation. *IEEE transactions on neural networks and learning systems*, 32(4):1654–1667, 2020.
- Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. URL <https://arxiv.org/abs/2109.05014>.

- Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.
- Desen Yuan. Language bias in visual question answering: A survey and taxonomy. *arXiv preprint arXiv:2111.08531*, 2021.
- Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv Preprint 2205.10625*, 2022. URL <https://arxiv.org/abs/2205.10625>.

## A APPENDIX

### A.1 BROADER IMPACT STATEMENT

We acknowledge that while the Img2Prompt 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 Img2Prompt. Future work could assess the magnitude of this bias and mitigate its impact.

### A.2 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

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

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 (Bian et al., 2021), 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  $k$ <sup>th</sup> 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:

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

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

### A.3 EXPERIMENTAL RESULTS OF DIFFERENT LLMs

Because some classic LLMs, e.g., GPT-3 Brown et al. (2020a)/PaLM Chowdhery et al. (2022) are not available, we only conducted the experiment with some open-sourced LLMs, e.g., GPT-J Wang & Komatsuzaki (2021), GPT-NEO Black et al. (2021) and BLOOM Scao et al. (2022). The results are shown as Table 7. The experimental results indicate that Img2Prompt enables various LLMs to perform VQA tasks, and that all of them are capable of achieving superior performance to zero-shot PICa and Frozen. This is a strong support for showing our method’s generalization ability with different LLMs. In addition, we find two interesting phenomena. (a) Other open-source LLMs perform marginally worse than OPT. This is due to the fact that other LLMs were trained on smaller datasets than OPT. (b) The multi-linguistic language model, BLOOM, performs marginally worse than English-specific LLMs.

Table 7: The experimental comparisons with different LLMs.

Methods	Shot Number	VQAv2 val	OK-VQA val
PICa <sub>GPT-3 175B</sub>	0	-	17.7
Ours <sub>GPT-Neo 2.7B</sub>	0	50.1	31.5
Ours <sub>OPT 2.7B</sub>	0	53.6	33.3
Ours <sub>GPT-J 6B</sub>	0	56.4	37.4
Ours <sub>BLOOM 7.1B</sub>	0	52.4	32.4
Ours <sub>OPT 6.7B</sub>	0	57.6	38.2

### 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 (Schwenk et al., 2022) as Table 10 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.

Table 8: The experimental comparisons with models trained in A-OKVQA training dataset.

Methods	A-OKVQA	
	Val	Test
<i>Models Fine-Tuned in A-OKVQA Training Set</i>		
Pythia (Jiang et al., 2018)	25.2	21.9
ViLBERT (Lu et al., 2019)	30.6	25.9
LXMERT (Tan & Bansal, 2019)	30.7	25.9
KRISP (Marino et al., 2021)	33.7	27.1
GPV-2 (Kamath et al., 2022)	<b>48.6</b>	<b>40.7</b>
<i>Zero-Shot Evaluation with Plug-in Frozen Large Language Model</i>		
Ours <sub>6.7B</sub>	33.3	32.2
Ours <sub>13B</sub>	33.3	33.0
Ours <sub>30B</sub>	36.9	36.0
Ours <sub>66B</sub>	38.7	38.2
Ours <sub>175B</sub>	42.9	<b>40.7</b>

### A.5 TEMPLATE-BASED QUESTION DESIGN

We design question templates for each part of speech type of answers as Table 9 shows.

### A.6 SENSITIVE ANALYSIS

We evaluate the sensitive analysis about the QA pairs and number of captions in prompt for LLM as Table 10 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 method is robust with different numbers of QA pairs and captions.

Table 9: 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 10: 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
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

Table 11: 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 12: 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**

(a)

**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 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 5: 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

(a)

**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 6: 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**

(a)

**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 7: 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

(a)

**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 8: 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**

(a)

**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 9: 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

(a)

**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 10: Failure case analysis for VQAv2. Red color indicates incorrect prediction.