From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models

Jiaxian Guo[1*]   Junnan Li[2]   Dongxu Li[2]   Anthony Meng Huat Tiong[2,3]
Boyang Li[3]   Dacheng Tao[1]   Steven Hoi[2]

[1] The University of Sydney   [2] Salesforce Research   [3] Nanyang Technological University

jguo5934@uni.sydney.edu.au   {junnan.li,li.d,anthony.tiong,shoi}@salesforce.com
boyang.li@ntu.edu.sg   dacheng.tao@gmail.com

Abstract

Large language models (LLMs) have demonstrated excellent zero-shot generalization to new language tasks. However, effective utilization of LLMs for zero-shot visual question-answering (VQA) remains challenging, primarily due to the modality disconnect and task disconnect between the LLM and VQA tasks. End-to-end training on multimodal data may bridge the disconnects, but is inflexible and computationally expensive. To address this issue, we propose Img2LLM, a plug-and-play module that provides LLM prompts to enable LLMs to perform zero-shot VQA tasks without end-to-end training. We develop LLM-agnostic models describe image content as exemplar question-answer pairs, which prove to be effective LLM prompts. Img2LLM offers the following benefits: 1) It achieves comparable or better performance than methods relying on end-to-end training. For example, we outperform Flamingo [3] by 5.6% on VQAv2. On the challenging A-OKVQA dataset, our method outperforms few-shot methods by as much as 20%. 2) It flexibly interfaces with a wide range of LLMs to perform VQA. 3) It eliminates the need to specialize LLMs using end-to-end finetuning and serve highly specialized LLMs to end users, thereby reducing cost. Code is available via the LAVIS [28] framework at https://github.com/salesforce/LAVIS/tree/main/projects/img2llm-vqa.

1. Introduction

Visual question answering (VQA) [5] 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 [5, 17] and others on logical reasoning [39]. However, human annotations are expensive to obtain and may introduce a variety of human biases [6, 10, 63], making the VQA system brittle towards new answer styles and question types [1, 21]. This has led researchers to zero-shot VQA methods [6, 10, 21] that do not require ground-truth question-answer annotations, thereby facilitating more generalizable VQA systems.

Recently, large language models (LLMs) (e.g., [8, 66]) have demonstrated excellent capabilities to perform tasks with zero in-domain data, conduct logical reasoning, and apply commonsense knowledge in NLP tasks [26, 55, 57]. As a result, recent approaches [3, 52, 61] have resorted to leverage LLMs in zero-shot VQA.

However, applying LLMs to VQA tasks is less than straightforward, due to (1) the modality disconnect between vision and language and (2) the task disconnect between language modeling and question answering. A common technique is to finetune a vision encoder jointly with the LLM [3, 20, 52] to align the vision and language representation spaces, but this can incur prohibitive computational and data cost. For example, Flamingo [3] finetunes on billions of image-text pairs with thousands of TPUs. Further, the finetuning specializes and introduces strong interdependence between the vision encoder and the LLM. If we need to upgrade the LLM as new versions emerge, the entire model needs to undergo expensive re-training.

In contrast to the end-to-end integration of LLM into a VQA system, this paper proposes a modular VQA system built on top of frozen off-the-shelf LLMs. This brings two benefits. First, it can reduce the deployment cost and simplify the deployment. Second, upgrading the LLM is straightforward. However, it is challenging to bridge the modality disconnect and task disconnect without end-to-end training. PICa [61] converts images into captions, and provides exemplar QA pairs from training data as prompt to the LLM. However, doing so assumes the existence of an-

---

*Work done while Jiaxian Guo was an intern at Salesforce Research.

notated training data and the performance is sensitive to the selection of few-shot exemplars.

We propose *Img2LLM*, a plug-and-play module that enables off-the-shelf LLMs to perform zero-shot VQA. The central insight of Img2LLM is that we can utilize a vision-language model (*e.g.* BLIP [30]) and a question-generation model to translate the image content into synthetic question-answer (QA) pairs, which are fed to the LLM as part of the prompt. These exemplar QA pairs tackle the modality disconnect by describing the image content verbally, and tackle the task disconnect by demonstrating the QA task to the LLM. Notably, the exemplar QA pairs are constructed entirely based on the test image and question, obviating the need for similar few-shot examples as required by PICa [61], which are not always available in practical zero-shot scenarios. When applied to the open-source OPT language models [66], Img2LLM achieves comparable or superior zero-shot VQA performance to methods that perform costly end-to-end training.

With this paper, we make the following contributions.

- We propose Img2LLM, a plug-and-play module that converts an image into synthetic question-answer pairs based solely on the current image of the question. Img2LLM bridges the modality disconnect between language and vision as well as the task disconnect between language modeling and visual question-answering.

- Img2LLM enables off-the-shelf LLMs to perform zero-shot VQA without costly end-to-end training or specialized textual QA networks [40], thereby allowing low-cost and flexible model deployment and painless LLM upgrades (Table 3).

- Our experimental results show that the OPT models equipped with Img2LLM achieve zero-shot VQA performance that is competitive or superior to the end-to-end trained models. For example, we outperform Flamingo [3] by 5.6% on VQAv2. We even outperform many 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, has been the focus of active research [2, 4, 5, 47, 62]. The past few years witnessed rapid performance advances with large-scale image-text pretraining [13, 19, 20, 30–32, 34, 49, 54, 64, 65] followed by fine-tuning on VQA datasets. To tackle knowledge-based VQA [39, 47], recent works [16, 18, 29, 33, 36–38, 59] incorporate external knowledge,

such as ConceptNet [50] or Wikipedia, but experimental results in [47] show that these methods still struggle to answer questions requiring complex reasoning.

2.2. LLM for Zero/Few-Shot VQA Tasks

Large language models (LLMs) [9, 12, 66] trained on web-scale corpus are powerful in natural language understanding and reasoning [8, 67]. 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 θ the model parameters. Prior VQA methods using LLMs mainly fall into two categories: multi-modal pretraining and language-mediated VQA.

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 finetune efficiently, [52] opt to fine-tune only the visual encoder while Flamingo [3] 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 [3].

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 [61] 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, [40] 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 [24], 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 [8] or denoising objectives [14] on language modeling tasks. As the LLMs are
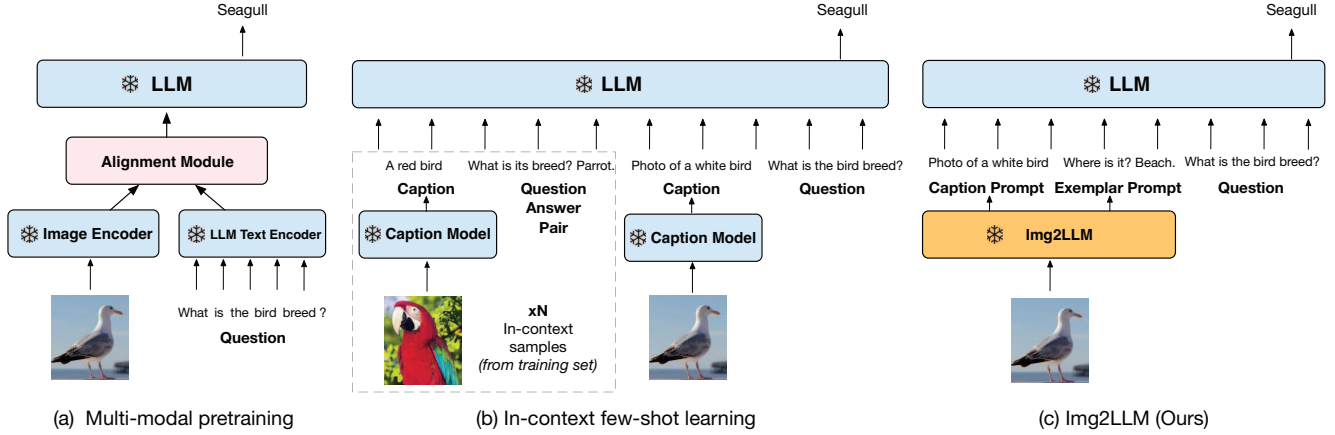
10868

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

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 [40, 61], 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 [61] or an LLM directly finetuned on textual QA [40]. 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, Img2LLM (Figure 1c), which generates image-relevant exemplar prompts for the LLM. Given a question Q and an image, our key insight is that we can generate synthetic question-answer pairs as in-context exemplars from the *current* image. The exemplars not only demonstrate the QA task but also communicate the content of the image to the LLM for answering the question Q, thereby hitting two birds with one stone. Img2LLM 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.

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 [10,27], we extract noun phrases (including named entities), verb phrases, adjective phrases, numbers, and boolean-typed words like "yes" and "no" as potential answers[1]. We show some extracted answer candidates in Figure 2 and Appendix A.3.

---

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

3.2. Question Generation

With the extracted answer candidate set $\{\hat{a}_j\}_{j=1}^U$, we can directly use any question generation network [2, 22, 25, 35, 60] to generate specific questions for each answer candidate. In this paper, we experiment with both template-based and neural question-generation methods. Note that to avoid violating the zero-shot requirements, our method is purely textual-based without access to any VQA data.

Template-based Question Generation. Using an off-the-shelf parser, we obtain the part-of-speech for each answer, and design specific question templates for each POS 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 put the complete list of templates in Appendix A.5.

Neural Question Generation. Inspired by [10], we train a neural question generation model on textual QA datasets. Specifically, we finetune a pretrained T5-large model [43] 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 [44], MultiRC [23], BookQA [41], CommonsenseQA [51] and Social IQA [45].

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 [8], which guides the LLM to perform QA task given the image content and bridges the task disconnect between language modelling and VQA.
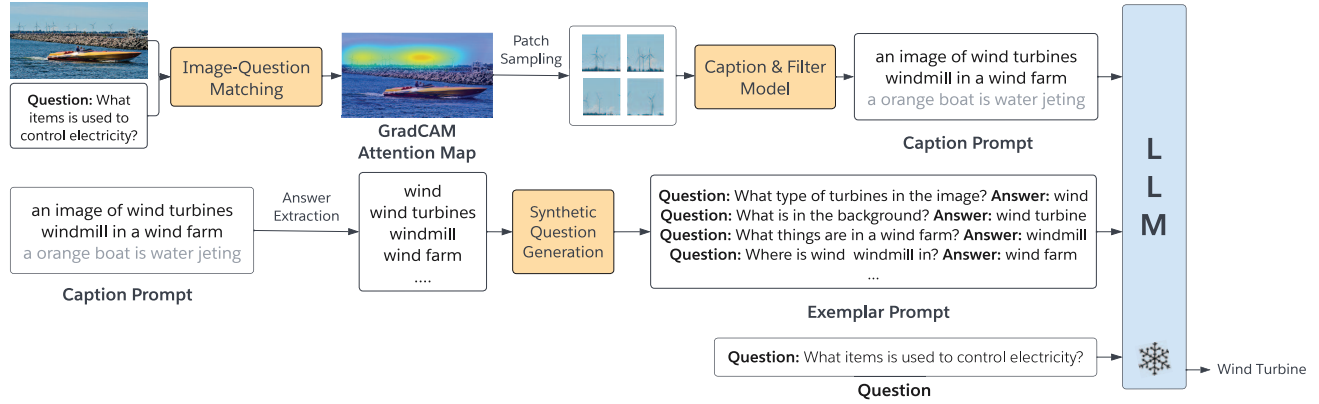
Figure 2. The overall pipeline of Img2LLM, including Caption Prompt and Exemplar Prompt generation.

Table 1. Results from mixing captions and exemplar prompts on 30B OPT [66].

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

As a sneak preview, we show effects of exemplar QA pairs in Table 1. The details of the instructions are explained in §3.4. 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.

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 [58] but generic captions generated by existing networks may not contain relevant information. In Figure 2, 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 whole 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 [30], as which assigns a similarity score $sim(v, q)$ to any pair of image v

and textual question q. With ITE, we use GradCAM [48], a feature-attribution interpretability technique, to generate a coarse localisation map highlighting matching image regions given a question [30]. Briefly, GradCam qualifies the cross-attention scores from the Transformer network by the gradient of ITE simlarity function $sim(v, q)$ with respect to the cross-attention scores. As this technique was proposed in [40], we leave the details to Appendix A.1.

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 [15]. To generate semantically meaningful captions, a short prompt, "a picture of," is also fed into the text decoder. We repeat this 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 exem-

plars. 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 §4.5 for analysis of caption selection strategies.

4. Experiment

In this section, we first validate the efficacy of Img2LLM 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 [17], OK-VQA [39] and A-OKVQA [47] datasets, which contain questions requiring perception, reasoning and commonsense to answer. Specifically, VQAv2 [17] contains 214,354 questions in the validation set and 107,394 in the test-dev dataset. OK-VQA [39] and A-OK-VQA [47] emphasize on commonsense reasoning, among which OK-VQA contains 5,046 test questions and A-OKVQA [47] contains 1,100 validation questions and 6,700 test questions.

Implementation details. To obtain question-relevant caption prompt, we use BLIP [30] 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. For the LLMs, our main result uses the open-source OPT model with multiple different sizes. Our ablation study also experiments with various other LLMs to show the generalization ability of our method. We use LLMs to generate answers auto-regressively, 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 rougly fall into three categories: (i) *Zero-shot methods with frozen LLMs*, such as PICa [61]. Our method also belongs to this category, yet unlike PICa, Img2LLM requires no training samples to compose the prompts. (ii) *Zero-shot methods with extra multi-modal pre-training*, such as Flamingo [3], Frozen [52], VL-T5 [11], FewVLM [20] and VLKD [13]. These methods require large-scale vision-language datasets and are costly to update. We also include results from VQ²A [10] and WeaQA [6] 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 [61], FewVLM [20] and ClipCap [42].

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 frozen LLMs. Img2LLM surpasses PICa, the best prior zero-shot model with frozen LLMs, by a significant margin (45.6 *versus* 17.7 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 across datasets. This shows that stronger language modelling capabilities help better comprehend the question, thus giving more accurate answers. Such a trend is 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.

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 [56] 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. Img2LLM obtains superior performance to most models with end-to-end pretraining, as

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 †; methods that answer from a predefined list of candidates are in grey. Further, ✗ annotates methods requiring no end-to-end training, which is desirable, and ✓ otherwise.

Methods	End-to-End Training?	Shot Number	VQAv2 val	VQAv2 test	OK-VQA test	A-OKVQA val	A-OKVQA test
Zero-Shot Evaluation with Frozen Large Language Model							
$PICa_{175B}$†	✗	0	-	-	17.7	-	-
$Img2LLM_{6.7B}$	✗	0	57.6	57.0	38.2	33.3	32.2
$Img2LLM_{13B}$	✗	0	57.1	57.3	39.9	33.3	33.0
$Img2LLM_{30B}$	✗	0	59.5	60.4	41.8	36.9	36.0
$Img2LLM_{66B}$	✗	0	59.9	60.3	43.2	38.7	38.2
$Img2LLM_{175B}$	✗	0	**60.6**	**61.9**	45.6	**42.9**	**40.7**
Zero-Shot Evaluation with Extra End-to-End Training							
$VL\text{-}T5_{no\text{-}vqa}$	✓	0	13.5	-	5.8	-	-
$FewVLM_{base}$	✓	0	43.4	-	11.6	-	-
$FewVLM_{large}$	✓	0	47.7	-	16.5	-	-
$VLKD_{ViT\text{-}B/16}$	✓	0	38.6	39.7	10.5	-	-
$VLKD_{ViT\text{-}L/14}$	✓	0	42.6	44.5	13.3	-	-
$Frozen_{7B}$	✓	0	29.5	-	5.9	-	
$Flamingo_{3B}$	✓	0	-	49.2	41.2	-	-
$Flamingo_{9B}$	✓	0	-	51.8	44.7	-	-
$Flamingo_{80B}$	✓	0	-	56.3	**50.6**	-	-
Zero-shot Evaluation with Access to Answer Candidates							
WeaQA ZSL	✓	0	46.8	-	-	-	-
VQ^2A	✓	0	61.1	-	19.8	-	-
Few-Shot Evaluation							
ClipCap→Cap→GPT_{175B}	✗	10	-	-	-	16.6	15.8
ClipCap→Rel→GPT_{175B}	✗	10	-	-	-	18.1	15.8
$FewVLM_{base}$	✓	16	48.2	-	15.0	-	
$FewVLM_{large}$	✓	16	51.1	-	23.1	-	-
$PICa_{175B}$†	✗	1	-	-	36.4	-	-
$PICa_{175B}$†	✗	4	-	-	43.3	-	-
$PICa_{175B}$†	✗	16	54.3	-	46.5	-	-
$PICa_{175B}$-Ensemble	✗	80	56.1	-	48.0	-	-

well as those evaluated in few-shot setups. For example, on VQAv2 our method surpasses $Flamingo_{80B}$, which cost over 500K TPU hours and billion-scale datasets to train, by a margin of 5.6 points. On A-OKVQA, Img2LLM more than doubles the best reported results so far, from Clip-Clap. The only a few exceptions are on OK-VQA, where our method obtains better results than $Flamingo_{9B}$, yet is not able to stay on par with $Flamingo_{80B}$. Considering that Img2LLM 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. Img2LLM achieves better performance than most supervised models, despite the fact that it uses zero training data and is evaluated in a zero-shot setup. These results once again validates its effectiveness.

Table 3. Zero-shot VQA performance with different LLMs.

Methods	VQAv2 val	OK-VQA
$PICa_{GPT\text{-}3\ 175B}$	-	17.7
$Frozen_{7B}$	29.5	5.9
$Ours_{GPT\text{-}Neo\ 2.7B}$	50.1	31.5
$Ours_{BLOOM\ 7.1B}$	52.4	32.4
$Ours_{GPT\text{-}J\ 6B}$	56.4	37.4
$Ours_{OPT\ 6.7B}$	57.6	38.2
$Ours_{OPT\ 175B}$	60.6	45.6

4.3. Experimental Results of Different LLMs

In Table 3, we evaluate the performance of Img2LLM on various open-sourced LLMs other than OPT, including GPT-J [53], GPT-Neo [7] and BLOOM [46]. The experimental results show that Img2LLM enables various LLMs to perform zero-shot VQA tasks, and that all of them achieve superior performance to zero-shot PICa [61] and Frozen [52]. This is a strong evidence for showing our

Table 4. Effect of question selection strategies.

		OK-VQA	VQAv2
PICa$_{175B}$		17.7	-
Agnostic	Random	35.9	52.9
Template	Random	40.2	53.0
	Max Freq.	41.5	55.8
Neural	Random	40.5	57.0
	Max Freq.	**41.8**	**59.5**

method's generalization ability with different LLMs.

4.4. Analysis on Question Generation Methods

Table 4 shows the performance of different question selection strategies described in Section 3.2. We compare three question generation techniques, include image-*agnostic*, which uses questions sampled 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 or important 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 7 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.

4.5. Ablation on Caption Selection

As Table 6 shows, we evaluate the performance different caption selection strategies, where Max Frequency se-

Table 5. Ablations on prompts designs.

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

Table 6. Ablation on caption selection methods.

Caption Selection	Random	Max Frequency	Min Frequency
OK-VQA Acc	41.3	41.1	**41.8**

lects 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.

4.6. Ablation Study on Prompt Design

We have two options to construct LLM's prompt. The first option is to append a syntheic 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-QAQAQA. 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.

4.7. Examples and Failure Case Analysis

In Figure 3, we show four examples of caption and exemplar prompts and the predictions, including cases of success and failure. In Figure 3(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 3(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.

5. Limitation

One limitation of the proposed approach is that generating image captions and question-answer pairs incurs ex-

Table 7. 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	**0.581**	58.1	0.0303	**0.821**
Question-relevant Caption with Filter	**41.8**	**0.025**	0.566	**59.5**	**0.0313**	0.804

Question: What type of profession is the man in red in?
GT Answer: bartender

Captions 1: a man in red shirt at a bar making drinks
Captions 2: a man in a red shirt is making a wine tasting
Captions 3: a man in a red shirt at a bar serving a bar

Synthetic Question 1: who is pouring a drink at a bar?
Answer: A man
Synthetic Question 2: where is a man in a red shirt making drinks? **Answer:** A bar
Question: What type of profession is the man in red in?
Predicted Answer: bartender

(a)

Question: The girl behind the man likely is of what relation to him?
GT Answer: daughter

Captions 1: a man is riding the back of a little girl on a motorcycle
Captions 2: an image of bearded man and a girl on a motorcycle riding on the motorcycle
Captions 3: man and child sitting on a motorcycle on the street

Synthetic Question 1: who is holding on to the bearded man on the back of the motorcycle?
Answer: A girl
Synthetic Question 2: what is the size of the girl riding on the motorcycle?
Answer: little
Question: The girl behind the man likely is of what relation to him?
Predicted Answer: daughter

(b)

Question: Why is he using knee pads?
GT Answer: Protection/Safety/Prevent injury

Caption 1: a skateboarder wearing knee pads on and protective gear on his knee
Caption 2: a man on skateboard in a helmet and knee pads
Caption 3: a skateboarder skateboarding with knee guards on

Synthetic Question 1: On what part of the body is a skateboarder wearing knee pads? **Answer:** Knee
Synthetic Question 2: What is the purpose of knee pads?
Answer: Protective
Question: Why is he using knee pads?
Predicted Answer: protect his knee

(c)

Question:what is the purpose of the wide tires on that bike?
GT answer:balance/traction/brake

Caption 1: a cargo bike sitting on a tire wheel.
Caption 2: the man is riding a bike on sands.
Caption 3: a man stands on a wheel on some sands.

Synthetic question 1:what are the tires on?
Answer: wheels
Synthetic question 2:what is a man doing on a bike?
Answer: riding
Question: What is the purpose of the wide tires on that bike?
Predicted answer: ride sand

(d)

Figure 3. Example predictions made by Img2LLM. Specifically, (a) and (b) are successful cases, while (c) and (d) are failure cases. See more examples at Appendix A.5.

tra inference overhead. On an 8×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. Notably, our method avoids expensive end-to-end multimodal representation alignment, which took more than 500K TPU hours in the case of Flamingo.

6. Conclusion

In this paper, we propose Img2LLM, 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, Img2LLM provides visual information and task guidance to LLMs in the format of easily-digestible prompts. This eliminates the require-ment for the expensive end-to-end vision-language alignment, increasing model deployment flexibility while decreasing model deployment cost. The experiments show that Img2LLM enables different LLMs to achieve comparable or even superior zero-shot VQA performance to other methods that require costly end-to-end training.

References

[1] 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*, pages 4971–4980, 2018. 1

[2] 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*, pages 2148–2166, 2021. 2, 3

[3] 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. 1, 2, 5

[4] 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*, pages 6077–6086, 2018. 2

[5] 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*, pages 2425–2433, 2015. 1, 2

[6] 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*, pages 3420–3435, Online, Aug. 2021. Association for Computational Linguistics. 1, 5

[7] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, Mar. 2021. If you use this software, please cite it using these metadata. 6

[8] 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, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1, 2, 3

[9] 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, 2020. 2

[10] 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. 1, 3, 5

[11] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 18–24 Jul 2021. 5

[12] 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. 2

[13] 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*, pages 2383–2395, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2, 5

[14] 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[15] 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)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. 4

[16] 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*, pages 489–498, 2020. 2

[17] 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)*, pages 6904–6913, 2017. 1, 5

[18] 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. 2

[19] 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*, pages 15513–15523, 2022. 2

[20] 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)*, pages 2763–2775, Dublin, Ireland, May 2022. Association for Computational Linguistics. 1, 2, 5

[21] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017. 1

[22] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, 2017. 3

[23] 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)*, pages 252–262, 2018. 3

[24] 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. 2

[25] 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. 3

[26] 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. 1

[27] 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. 3

[28] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. Lavis: A library for language-vision intelligence. *arXiv preprint arXiv:2209.09019*, 2022. 1

[29] 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*, pages 1227–1235, 2020. 2

[30] 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. 2, 4, 5

[31] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc., 2021. 2

[32] 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, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 2

[33] 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. 2

[34] 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

[35] 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*, pages 10437–10446, 2020. 3

[36] 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. 2

[37] 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. 2

[38] 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

[39] 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*, pages 3195–3204, 2019. 1, 2, 5

[40] 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. In *EMNLP Findings*, 2022. 2, 3, 4

[41] 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. 3

[42] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5

[43] 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. 3

[44] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018. 3

[45] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019. 3

[46] 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. 6

[47] 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. 2, 5

[48] 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)*, pages 618–626, 2017. 4

[49] 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. 2

[50] 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. 2

[51] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018. 3

[52] 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, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc., 2021. 1, 2, 5, 6

[53] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021. 6

[54] 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. 2

[55] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma,

[56] 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*, 2022. 5

[57] 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*, 2022. 1

[58] Jialin Wu, Zeyuan Hu, and Raymond J Mooney. Generating question relevant captions to aid visual question answering. *arXiv preprint arXiv:1906.00513*, 2019. 4

[59] 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, pages 2712–2721, 2022. 2

[60] 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. 3

[61] 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. 1, 2, 3, 5, 6

[62] 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*, pages 21–29, 2016. 2

[63] Desen Yuan. Language bias in visual question answering: A survey and taxonomy. *arXiv preprint arXiv:2111.08531*, 2021. 1

[64] 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. 2

[65] 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*, pages 5579–5588, 2021. 2

[66] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui 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. 1, 2, 4

[67] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv Preprint 2205.10625*, 2022. 2