Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training

Anthony Meng Huat Tiong[1,2], **Junnan Li**[1], **Boyang Li**[2],
Silvio Savarese[1], and **Steven C.H. Hoi**[1]
[1]Salesforce Research [2]Nanyang Technological University, Singapore
{anthony.tiong, junnan.li, ssavarese, shoi}@salesforce.com
boyang.li@ntu.edu.sg
https://github.com/salesforce/LAVIS/tree/main/projects/pnp-vqa

Abstract

Visual question answering (VQA) is a hallmark of vision and language reasoning and a challenging task under the zero-shot setting. We propose Plug-and-Play VQA (PNP-VQA), a modular framework for zero-shot VQA. In contrast to most existing works, which require substantial adaptation of pretrained language models (PLMs) for the vision modality, PNP-VQA requires no additional training of the PLMs. Instead, we propose to use natural language and network interpretation as an intermediate representation that glues pretrained models together. We first generate question-guided informative image captions, and pass the captions to a PLM as context for question answering. Surpassing end-to-end trained baselines, PNP-VQA achieves state-of-the-art results on zero-shot VQAv2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019). With 11B parameters, it outperforms the 80B-parameter Flamingo model (Alayrac et al., 2022) by 8.5% on VQAv2. With 738M PLM parameters, PNP-VQA achieves an improvement of 9.1% on GQA over FewVLM (Jin et al., 2022) with 740M PLM parameters.

1 Introduction

Recent years have witnessed unprecedented performance gains on many natural language reasoning tasks, especially in zero-shot and few-shot settings, being derived from scaling up pretrained language models (PLMs) and their training data (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Raffel et al., 2020; Black et al., 2022; Sanh et al., 2022; Wei et al., 2021). Inspired by their success, a natural thought is that utilizing PLMs should also boost zero-shot performance in vision-language reasoning tasks.

However, to leverage PLMs for vision-language tasks, most existing methods require non-trivial adaptation of the PLMs for the vision modality, which necessitates the design of new network components and training objectives. For example, Sung et al. (2022) and Alayrac et al. (2022) insert into the PLMs new layers that are trained from scratch. Tsimpoukelli et al. (2021) train vision encoders that output soft prompts to frozen PLMs. Chen et al. (2022) and Eichenberg et al. (2021) train both the vision encoders and new layers inserted into PLMs. In the zero-shot setting, various vision-language pretraining objectives are employed, such as image captioning (Alayrac et al., 2022) and image-conditioned masked language modeling (Jin et al., 2022).

From the perspective of general-purpose AI, the ability to perform new tasks by simply recombining large-scale pretrained models, or foundation models (Bommasani et al., 2021), without architectural changes or extra training would be highly desirable. Such a system would be able to dynamically adjust to previously unknown tasks by simply rewiring a small number of foundation models. However, to obtain high performance without some form of end-to-end training would seem difficult, if not impossible.

We present Plug-and-Play VQA (PNP-VQA), a framework for zero-shot visual question answering which conjoins large pretrained models with zero additional training and achieves state-of-the-art performance on zero-shot VQAv2 (Goyal et al., 2017) and GQA (Hudson and Manning, 2019). For the purpose of bridging the vision and language modalities, we employ a pretrained vision-language model (PVLM) (Li et al., 2022b) that describes visual information with textual captions. In order to obtain relevant and informative captions, we apply a network interpretability technique (Selvaraju et al., 2017) to detect image patches that are relevant to the question. After that, we generate captions stochastically for these image patches. Finally, we employ a PLM (Khashabi et al., 2022) to answer the question from the captions.

Research in cognitive science and neuroscience suggests that the human cognitive system is largely

modular (Shettleworth, 2012; Bertolero et al., 2015). For instance, the pioneering work of Fodor (1983) argued that the low-level human cognition is constituted of several fast, autonomous, and domain-specific modules. For purely practical purposes, a modular design of artificial general intelligence would make it easy to harness rapid progress in each individual component, as the components can be individually replaced and updated without affecting other parts of the system. With this paper, we offer such a modular design for zero-shot VQA that leverages recent advances in PLM and PVLMs and combines them with an innovative application of network interpretability.

We summarize our contributions as follows:

- We introduce PNP-VQA, a modular framework for zero-shot VQA without training. Its flexibility allows PNP-VQA to jointly evolve as pretrained models continue to advance.
- Besides natural language, we propose the use of network interpretation as the interface between pretrained LMs and VLMs. With an interpretability technique, we create image captions that extensively cover information relevant to the question, which enable accurate QA.
- We demonstrate state-of-the-art zero-shot VQA performance on multiple benchmarks. On VQAv2, PNP-VQA$_{11B}$ obtains 8.5% improvement over Flamingo$_{80B}$ (Alayrac et al., 2022), which applies extensive end-to-end VL-pretraining. On GQA, PNP-VQA$_{large}$ outperforms FewVLM$_{large}$ (Jin et al., 2022) by 9.1%.

2 Related Work

Large-scale image-text pretraining of neural networks is a popular research direction. Various vision-language pretraining tasks have been proposed, including image-conditioned language modeling (Tsimpoukelli et al., 2021; Alayrac et al., 2022), masked language modeling (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2021b), prefix language modeling (Wang et al., 2022), image-text matching (Li et al., 2019; Chen et al., 2020; Li et al., 2020) and image-text contrastive learning (Radford et al., 2021; Jia et al., 2021; Li et al., 2021a). After pretraining, several models exhibit zero-shot capabilities in image-text retrieval (Jia et al., 2021; Radford et al., 2021; Zeng et al., 2022b) and image captioning (Wang et al., 2022; Li et al., 2022b). However, zero-shot VQA remains a challenging task due to its high requirement on the model's reasoning ability.

Adapting PLMs for zero-shot VQA has shown promising results. In order to incorporate vision information into PLMs, most existing methods perform additional vision-language training on image-text data. Frozen (Tsimpoukelli et al., 2021) trains the vision encoder while keeping the gigantic PLM frozen to retain its knowledge in question answering. The output from the vision encoder is prepended to the text as prompts to the frozen language model. FewVLM (Jin et al., 2022) finetunes the PLM using the prefix language modeling and masked language modeling objectives. VLKD (Dai et al., 2022) distills multimodal knowledge to PLM by using CLIP (Radford et al., 2021) as the teacher model during finetuning. Flamingo (Alayrac et al., 2022) adds additional layers to both the pretrained vision model and the PLM and trains the new layers on billions of image-text pairs.

Different from the above work, PNP-VQA directly employs pretrained models with neither architectural modifications nor additional training.

Most similar to our work, PICa (Yang et al., 2022) converts an image to a single caption and adopts GPT-3 (Brown et al., 2020) for zero-shot VQA. In comparison, PNP-VQA generates multiple question-guided captions and performs fusion of captions after encoding to effectively utilize a large number of captions, yielding considerable performance gains.

An orthogonal research direction for zero-shot VQA is to train the VLMs on synthetic VQA examples generated from captions (Changpinyo et al., 2022; Banerjee et al., 2021). PNP-VQA does not require additional training.

Natural language as an intermediate representation or interface between different models or multiple steps of reasoning is an emerging machine learning strategy. It dates back to at least Andreas et al. (2018) and saw renewed interest in the past few months due to the prevalence of large PLMs. Andreas et al. (2018) and Vong and Lake (2022) learn natural language descriptions that function as few-shot classifiers within an image-text matching model. Bostrom et al. (2022) generate intermediate reasoning steps with finetuned PLMs. Zhou et al. (2022) prompt a PLM to generate subproblem descriptions for a complex problem, and feed the subproblems back to the PLM to solve hierarchically.

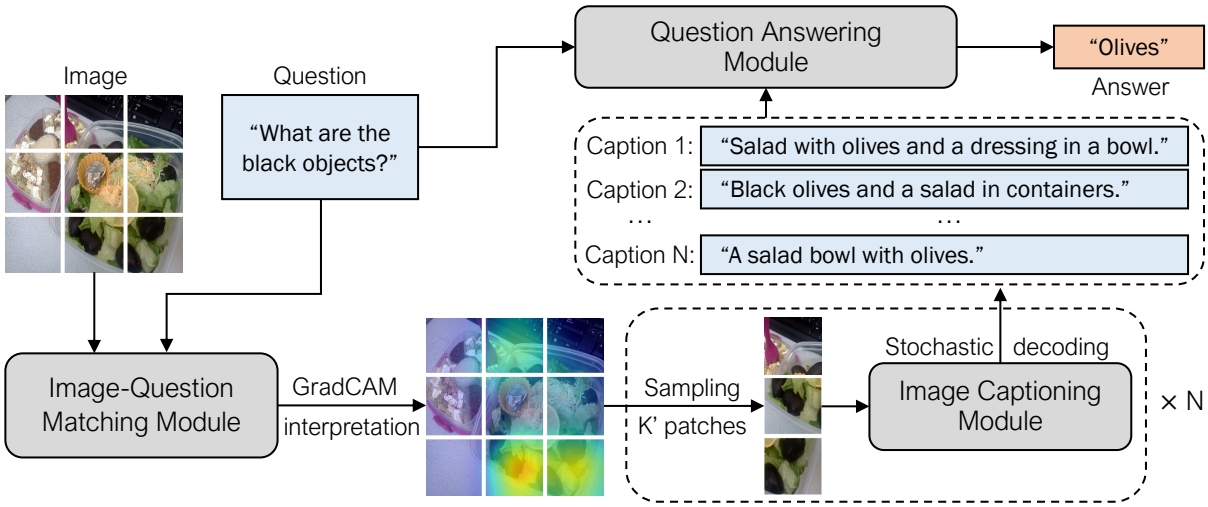


Figure 1: The system architecture of PNP-VQA, consisting of three pretrained modules: (1) an image-question matching module that identifies image patches relevant to the question, (2) an image captioning module that generates a diverse set of captions, (3) a question answering module that generates an answer given the question and captions. For the image-question matching module and image captioning module, we adopt BLIP (Li et al., 2022b). For the question answering module, we adopt UnifiedQAv2 (Khashabi et al., 2022).

Wu et al. (2022) chain PLM outputs and inputs. Zeng et al. (2022a) show that language-conjoined LM and VLM successfully perform captioning and retrieval but do not evaluate their models on VQA. In comparison, PNP-VQA adopts both natural language and network interpretation as the interface between different pretrained models.

3 Method

The central idea of Plug-and-Play VQA (PNP-VQA) is to establish an interface between a pretrained language model and a pretrained vision-language model without training. We demonstrate that natural language image captions and network saliency maps together serve as an effective interface. Ideally, the generated captions should thoroughly cover information that is present in the image and be relevant to the question. We foster relevance by identifying image patches most related to the question with a saliency map-based interpretability technique and generating captions from these patches only. Further, we promote coverage by injecting stochasticity, including random sampling of relevant image patches and of the textual tokens during caption generation.

The overall system architecture (Figure 1) consists of three modules:

1. an image-question matching module that identifies the relevant image patches given a question,
2. an image captioning module that generates a diverse set of captions from a set of image patches, and
3. a question answering module that outputs an answer given the question and the generated captions.

In this section, we introduce the three modules in detail.

3.1 Matching Image Patches and Questions

An image serves as a rich source of information, but the question at hand is likely focused only on particular objects or regions. Therefore, we encourage PNP-VQA to generate captions that describe image regions relevant to the question instead of generic captions with no specific aim.

We accomplish this goal by leveraging BLIP (Li et al., 2022b), a large-scale pretrained vision-language model that contains a network branch outputting a similarity score $\text{sim}(v, t)$ between an image v and a text t. This branch, called Image-grounded Text Encoder (ITE), employs a vision transformer (Dosovitskiy et al., 2021) that encodes the image, and a textual encoder that attends to the image features using cross-attention. As input to the image encoder, the image is equally divided into K patches.

To identify relevant image patches, we feed the image v and the question t to the ITE network and apply a variation of GradCAM (Selvaraju et al.,

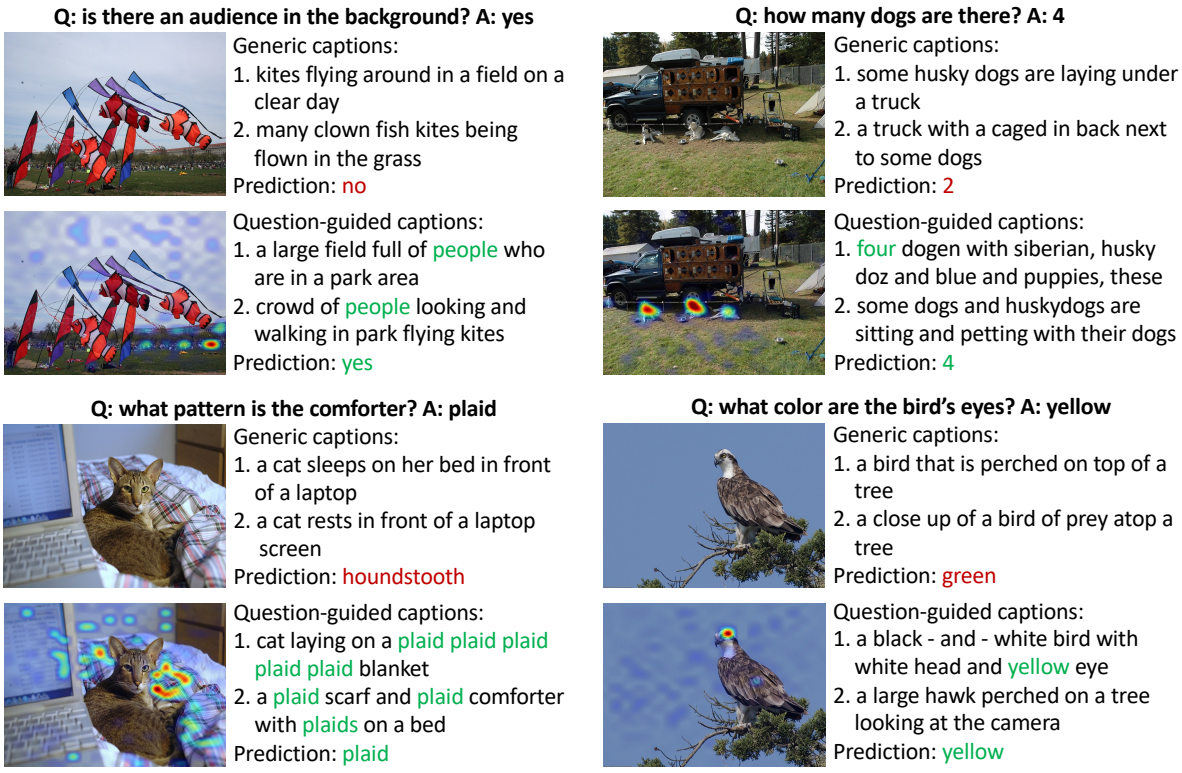


Figure 2: Examples of generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps on VQAv2 data. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues from captions. Words in red indicate wrong answer predictions.

2017), a feature-attribution interpretability technique, that aggregates all cross-attention maps using weights from the gradients. Formally, let us denote image patch features as $X \in \mathbb{R}^{K \times D_v}$, where K is the number of image patches and D_v the image feature dimension. We denote textual features as $Y \in \mathbb{R}^{M \times D_t}$, where M is the number of textual tokens and D_t the text feature dimension. For every cross-attention head, we have parameter matrices $W_Q \in \mathbb{R}^{D_t \times D_t}$ and $W_K \in \mathbb{R}^{D_v \times D_t}$. The cross-attention scores, $A \in \mathbb{R}^{M \times K}$, can be written as

$$A = \text{softmax}\left(\frac{YW_Q W_K^\top X^\top}{\sqrt{D_t}}\right). \quad (1)$$

The j^{th} row of A indicates the amount of attention the j^{th} textual token allocates to all image patches. At a selected layer of the ITE network, we compute the derivative of the similarity score w.r.t the cross-attention score, $\partial\ \text{sim}(v,t)/\partial A$, and multiply the gradient matrix element-wise with the cross-attention scores. The relevance of the i^{th} image patch, $\text{rel}(i)$, takes the average over H attention heads and the sum over M textual tokens:

$$\text{rel}(i) = \frac{1}{H}\sum_{j=1}^{M}\sum_{h=1}^{H} \max\left(0, \frac{\partial\ \text{sim}(v,t)}{\partial A_{ji}^{(h)}}\right) A_{ji}^{(h)}, \quad (2)$$

where the superscript $^{(h)}$ denotes the index of attention heads. For every caption we generate, we sample a subset of K' image patches with probability proportional to the patch relevance. The captioning module sees the sampled patches only.

We provide the following motivation for the technique. The attention matrix A may be taken as indicative of patch importance. However, much redundancy exists among these matrices and many attention heads may be pruned with little performance loss (Bian et al., 2021), suggesting that some scores are uninformative. Inspired by GradCAM, we filter out uninformative attention scores by multiplication with the gradient which could cause an increase in the image-text similarity.

Figure 2 shows some examples of generic captions and question-guided captions with associated relevance heatmaps. We can clearly observe that question-guided captions contain more relevant information that helps produce the correct answers.

Image Patch Sampling Strategy	Num. of Captions	VQAv2	OK-VQA	GQA
No captions	0	33.4	10.3	25.9
All patches (generic captions)	5	53.5	26.6	36.5
Uniform random sampling	5	52.0	25.5	36.2
Question-guided patch sampling	5	56.3	27.0	37.9
Human-written captions from MS COCO	5	56.9	28.1	-
All patches (generic captions)	100	58.6	31.9	39.8
Uniform random sampling	100	58.4	32.4	40.4
Question-guided patch sampling	100	**62.1**	**34.1**	**42.3**

Table 1: Comparison of different sampling strategies for image patches. 100 question-guided captions surpass the performance of 5 human-written captions from MS COCO.

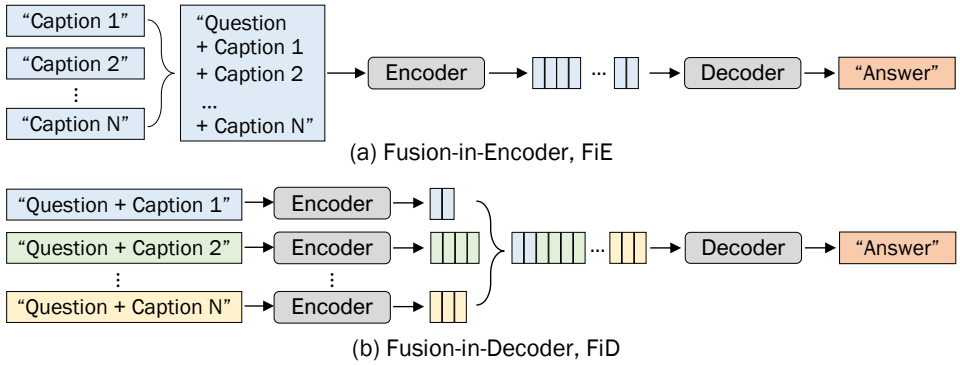


Figure 3: Two methods to process multiple captions with a question answering model. (a) Fusion-in-Encoder (FiE), which concatenates the captions as a long input paragraph to the encoder. (b) Fusion-in-Decoder (FiD), which encodes each caption with the question individually and concatenates all encoded representations as input to the cross-attention mechanism of the decoder.

Table 1 gives a quantitative analysis about the effect of different patch selection methods on zero-shot VQA performance across three datasets. Question-guided patch sampling substantially outperforms generic captioning using all patches and random patch sampling, especially when the number of captions is large. 100 question-guided captions outperform the 5 human-written captions from MS COCO by 5.2% on VQAv2 and 6.0% on OK-VQA, demonstrating the merit of the proposed approach.

3.2 Informative Image Captioning

Even with relevant image regions, there may still be more than one way to describe these regions. Some descriptions may contain the desired answer to the question, whereas others may not. Without the ability to identify the answer *a priori*, we aim to generate maximally diverse captions to provide coverage of possible answers.

We adopt the image captioning network branch from BLIP (Li et al., 2022b) and apply stochastic top-k sampling (Fan et al., 2018) instead of beam search, which is known to produce dull and repetitive captions (Vijayakumar et al., 2018; Holtzman et al., 2020). The input to the network contains the K' image patches sampled according to relevance (see §3.1). We prepend a short prompt, "a picture of " as input to the text decoder. We repeat this process to generate N captions per image to encourage diversity of captions and coverage of visual content. To prevent repetition, we keep a generated caption only if it is not subsumed by any previous caption as an exact substring.

3.3 Answering the Question

The question-answering encoder-decoder model is pretrained on text data only and can only process text. Therefore, we include the question and the generated captions as input to the model. As discussed in §3.2, the image captioning module generates multiple diverse captions. To process such long inputs efficiently, we adopt the Fusion-in-Decoder (FiD) strategy (Izacard and Grave, 2021).

We illustrate the FiD strategy in Figure 3 by comparing it with the more straightforward Fusion-in-Encoder (FiE), which concatenates the question and all captions into a long paragraph as input to

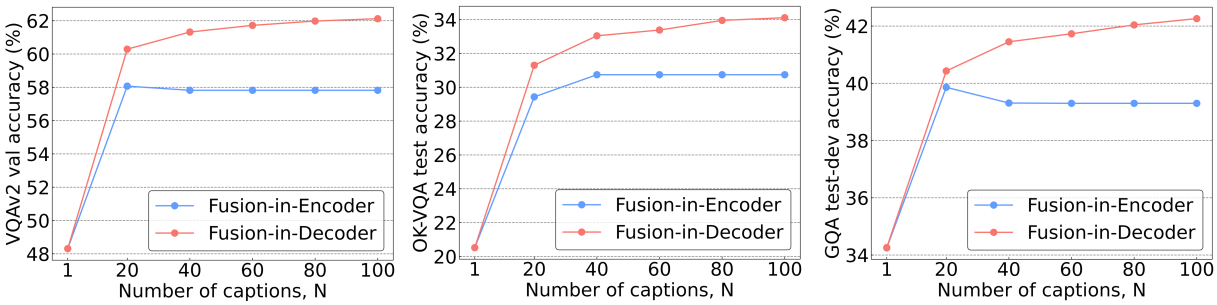


Figure 4: Comparison between Fusion-in-Encoder and Fusion-in-Decoder for VQAv2, OK-VQA and GQA.

the encoder. In contrast, FiD encodes each caption with the question separately and concatenates the *encoded representations* of all tokens from all captions. The result is fed as input to the decoder and is processed through the cross-attention mechanism. Since the time complexity of the self-attention mechanism scales quadratically with input length, whereas the cross-attention scales linearly with the encoder's output length, FiD is much more efficient than FiE. Further, FiE is constrained by the maximum input length of the encoder, caused by the positional encoding, but FiD does not have this constraint. Hence, with FiD, PNP-VQA can benefit from even more captions.

We plot the performance of FiD and FiE against the number of captions in Figure 4. Initially, both methods improve as the number of captions increases. However, the performance of FiE is capped at around 40 captions when the maximum input length is exceeded, whereas the performance of FiD continues to rise.

4 Experiments

4.1 Datasets and Evaluation

We adopt multiple zero-shot VQA benchmarks, including the validation set (214,354 questions) and test-dev set (107,394 questions) of VQAv2 (Goyal et al., 2017), the test set (5,046 questions) of OK-VQA (Marino et al., 2019), and the test-dev set (12,578 questions) of GQA-balanced (Hudson and Manning, 2019). We include the VQAv2 validation set as a few recent works (Tsimpoukelli et al., 2021; Jin et al., 2022) evaluate their performance on this dataset only. We obtain the answer by open-ended generation and perform evaluation based on exact matching. We report soft-accuracy (Goyal et al., 2017) for VQAv2 and OK-VQA to account for multiple ground truth answer; for GQA, we report the standard accuracy.

4.2 Implementation Details

To obtain the image-question matching module and image captioning module, we adopt BLIP (Li et al., 2022b) with the ViT-L/16 architecture pretrained on 129M image-text pairs. The original BLIP-ITM and BLIP-Caption models further finetune on the 2017 train split of COCO Captions (Lin et al., 2014), which partially overlaps with VQAv2 and OKVQA. To prevent data leak, we instead finetune on the 2014 train split of COCO Captions, which does not overlap with the VQA evaluation datasets. We emphasize that this represents less, not more, training compared to the publicly released BLIP.

For the question answering module, we adopt UnifiedQAv2 (Khashabi et al., 2022) trained on diverse textual QA datasets. It is worth noting that UnifiedQAv2 is completely unaware of the visual modality during training. Therefore, its training data do not overlap with the VQA datasets.

Unless otherwise stated, we utilize a total of 100 captions per question. We select the 8^{th} cross-attention layer of the ITE network for GradCAM. We sample $K' = 20$ image patches for the generation of each caption, and use $k = 50$ for top-k decoding (see Fig. 9 in Appendix B). For VQAv2 and OK-VQA, we apply FiD and encode the question with one caption at a time. However, for GQA, we encode each question with a group of 5 captions. GQA requires compositional visual reasoning and thus benefits from more contextual information per question. We perform experiments using LAVIS (Li et al., 2022a) on 8 Nvidia A100 GPUs.

4.3 Comparison with State of the Arts

We compare with state-of-the-art methods that formulate zero-shot VQA as open-ended answer generation. We categorize the methods based on how the pretrained networks are conjoined. In the first group, including VL-T5$_{\text{no-vqa}}$ (Cho et al., 2021), FewVLM (Jin et al., 2022), VLKD (Dai

Method	Language			Vision			VQAv2		OK-VQA	GQA
	Model	#Params	VL-aware	Model	#Params	VL-aware	Val	Test-dev	Test	Test-dev
Pretrained models conjoined by end-to-end VL training.										
VL-T5$_{\text{no-vqa}}$	T5	224M	✓	Faster R-CNN	64M	✗	13.5	-	5.8	6.3
FewVLM$_{\text{base}}$	T5	224M	✓	Faster R-CNN	64M	✗	43.4	-	11.6	27.0
FewVLM$_{\text{large}}$	T5	740M	✓	Faster R-CNN	64M	✗	47.7	-	16.5	29.3
VLKD$_{\text{ViT-B/16}}$	BART	407M	✓	ViT-B/16	87M	✓	38.6	39.7	10.5	-
VLKD$_{\text{ViT-L/14}}$	BART	408M	✓	ViT-L/14	305M	✓	42.6	44.5	13.3	-
Flamingo$_{\text{3B}}$	Chinchilla-like	2.6B	✓	NFNet-F6	629M	✓	-	49.2	41.2	-
Flamingo$_{\text{9B}}$	Chinchilla-like	8.7B	✓	NFNet-F6	629M	✓	-	51.8	<u>44.7</u>	-
Flamingo$_{\text{80B}}$	Chinchilla	80B	✓	NFNet-F6	629M	✓	-	56.3	**50.6**	-
Frozen	GPT-like	7B	✗	NF-ResNet-50	40M	✓	29.5	-	5.9	-
Pretrained models conjoined by natural language and zero training.										
PICa	GPT-3	175B	✗	VinVL-Caption	259M	✓	-	-	17.7	-
PNP-VQA$_{\text{base}}$	UnifiedQAv2	223M	✗	BLIP-Caption	446M	✓	54.3	55.2	23.0	34.6
PNP-VQA$_{\text{large}}$	UnifiedQAv2	738M	✗	BLIP-Caption	446M	✓	57.5	58.8	27.1	38.4
PNP-VQA$_{\text{3B}}$	UnifiedQAv2	2.9B	✗	BLIP-Caption	446M	✓	<u>62.1</u>	<u>63.5</u>	34.1	**42.3**
PNP-VQA$_{\text{11B}}$	UnifiedQAv2	11.3B	✗	BLIP-Caption	446M	✓	**63.3**	**64.8**	35.9	<u>41.9</u>

Table 2: Comparison with state-of-the-art models on zero-shot VQA. Flamingo (Alayrac et al., 2022) inserts additional parameters into the language model and perform training using billion-scale vision-language data. The best accuracy is bolded and the second best is underlined.

et al., 2022), Flamingo (Alayrac et al., 2022), and Frozen (Tsimpoukelli et al., 2021), a vision encoder (VE) embeds the image as a dense matrix and feeds it to the pretrained language model (PLM). After that, the system performs a round of end-to-end vision-language (VL) training on tasks other than VQA, such as image captioning. VL-T5$_{\text{no-vqa}}$ and FewVLM freeze the VE and finetune the PLM, whereas Frozen freezes the PLM and trains the VE. VLKD finetunes both the PLM and part of VE. Flamingo partially finetunes both the VE and the PLM. In the second group, the two foundation models are not jointly trained. Instead, they use language in the form of captions as the intermediate representation for an image. This group includes PICa (Yang et al., 2022) and our proposed model, PNP-VQA.

Table 2 shows the results. PNP-VQA outperforms previous methods by large margins on VQAv2 and GQA. On VQAv2 test-dev, PNP-VQA$_{\text{11B}}$ outperforms the second best technique, Flamingo$_{\text{80B}}$ (Alayrac et al., 2022), by 8.5%. PNP-VQA$_{\text{3B}}$ outperforms Flamingo$_{\text{80B}}$ by 7.2% despite its significantly smaller size and the similar-sized Flamingo$_{\text{3B}}$ by 14.3%. On GQA, PNP-VQA$_{\text{large}}$ outperforms the FewVLM$_{\text{large}}$ by 9.1%, with similar-sized PLM despite the lack of end-to-end training. Only on OK-VQA, Flamingo performs better than PNP-VQA. OK-VQA requires external knowledge not existing in the images and cannot be solved by good captions alone. We hypothesize that the end-to-end training on the gigantic vision-language dataset of Flamingo induces a mapping between images and knowledge concepts that helps with OK-VQA. However, PNP-VQA is still better on OK-VQA than all other baselines that not trained on the gigantic Flamingo data. Compared with language-conjoined PICa (Yang et al., 2022) with 175B parameters, PNP-VQA$_{\text{11B}}$ achieves a sizable improvement of 18.2%.

The results underscore the difficulty of zero-shot VQA using language models without any vision-language (VL) training. PICa, with its 175B-parameter language model, achieves comparable performance as FewVLM$_{\text{large}}$, whose language model is 236x smaller but finetuned on VL data. On the other hand, finetuning the billion-scale language model could incur heavy computational cost and risk catastrophic forgetting (Tsimpoukelli et al., 2021; Alayrac et al., 2022). PNP-VQA demonstrates the feasibility of a different paradigm: using billion-scale pretrained language models for VQA with zero training.

5 Analysis

5.1 Are PNP-VQA captions informative?

Intuitively, if the captions contain the correct answer, the QA model would have a higher chance to answer correctly. To measure the utility of captions, we compute the *answer hit rate* (AHR), or the proportion of questions for which at least one caption contains the ground-truth answer verbatim. Here we exclude questions with yes/no answers as the meaning of "yes" and "no" can be contextual

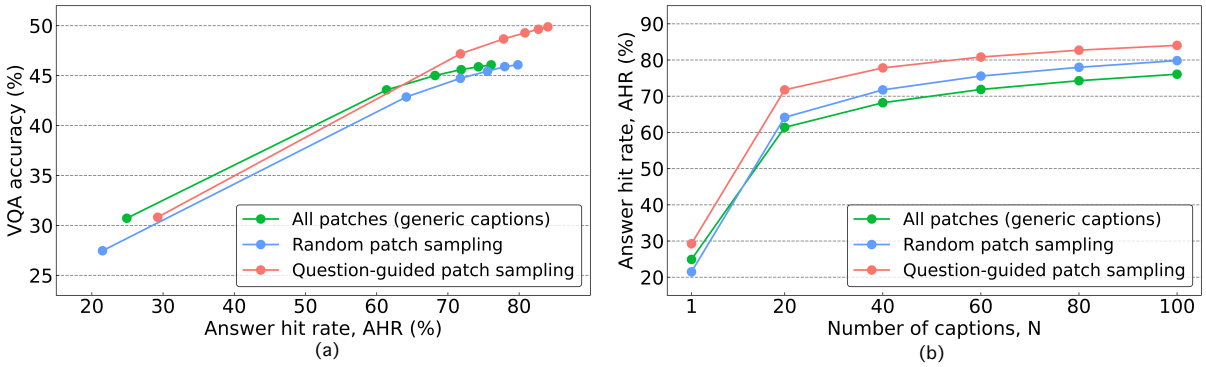


Figure 5: Analysis on the relationships between *answer hit rate* (AHR), VQA accuracy, and the number of captions per question (N). (a) shows a positive correlation between AHR and VQA accuracy. (b) shows the AHR increases with N, where the proposed question-guided patch sampling produces captions with the highest AHR.

Decoding Method	VQAv2	OK-VQA	GQA
Beam search	55.3	26.8	37.2
Temperature (t=0.5)	61.2	32.1	41.4
Temperature (t=1)	60.0	31.6	41.6
Nucleus (p=0.9)	61.3	32.9	41.7
Nucleus (p=0.95)	60.7	32.2	41.9
Top-k (k=50)	**62.1**	**34.1**	42.3
Top-k (k=100)	61.9	34.0	**42.3**

Table 3: Ablation study on different caption decoding methods. PNP-VQA$_{3B}$ performs well across the stochastic methods.

QA Model	#Params	VQAv2	OK-VQA	GQA
GPT-J	6B	28.7	14.5	18.5
T0	3B	49.6	26.6	32.3
T0	11B	47.3	30.5	33.4
UnifiedQAv2	3B	62.1	34.1	**42.3**
UnifiedQAv2	11B	**63.3**	**35.9**	41.9

Table 4: Ablation study on various textual question answering module for PNP-VQA on zero-shot VQA. UnifiedQAv2 is a task-specific model pretrained for question answering.

and these two words appear rarely in captions.

Figure 5(a) shows the correlation between the AHR and VQA accuracy, computed over the VQAv2 validation set, for three techniques of image patch sampling: question-guided sampling, uniform random sampling, and all patches. We observe that, within each sampling method, the VQA accuracy increases as the AHR increases. This corroborates our hypothesis that the presence of the answer in the captions facilitates the generation of the correct answer.

The correlation between performance and AHR is not perfect, as AHR does not capture other factors that may affect the answer accuracy, such as the position of the answer in the sentence and the number of its occurrence. However, AHR provides an easy-to-compute and useful measure for the information quality of the captions.

Figure 5(b) shows how AHR changes with the number of captions. Among the three techniques, question-guided sampling produces captions with the highest AHR. Thus, we may attribute the good performance of PNP-VQA partially to its informative, question-guided captions that directly contain the correct answer. Further, as the number of captions increases from 20 to 100, question-guided AHR increases from 71.8% to 84.0%. This demonstrates the benefit of Fusion-in-Decoder, which allows PNP-VQA to utilize up to 100 captions.

5.2 How sensitive is PNP-VQA to the caption decoding method?

As the content of captions plays a crucial role in the performance of PNP-VQA, we investigate the sensitivity to the choice of the caption decoding methods. We test four methods, including the deterministic beam search and three stochastic methods — temperature sampling (Ficler and Goldberg, 2017; Caccia et al., 2020), nucleus sampling (Holtzman et al., 2020), and top-k sampling (Fan et al., 2018). We generate 100 captions from each method, and report the results in Table 3. PNP-VQA performs very similarly across stochastic decoding methods, but beam search results in a noticeable drop. Upon close inspection, we observe that beam search generates repetitive captions that do not sufficiently cover different aspects of the image.

5.3 Can PNP-VQA work with other textual QA models?

We experiment with two other PLMs as the question answering module for PNP-VQA: T0 (Sanh et al., 2022) and GPT-J (Wang and Komatsuzaki, 2021). T0 is an encoder-decoder model which is pretrained in a multi-task fashion on a collection of NLP tasks, including question answering. GPT-J is a decoder-only model, a much smaller open-source alternative to GPT-3 (Brown et al., 2020), which is pretrained with a task-agnostic language modeling loss on a large-scale text corpus. Table 4 shows that UnifiedQAv2 performs better on VQA tasks compared to T0 and GPT-J. We attribute UnifiedQAv2's good performance to the fact that it is a task-specific question answering model with superior textual QA performance. The result indicates that the choice of PLM is important when performing zero-shot VQA with zero training. The modular and flexible design of PNP-VQA leaves room for further performance improvements as more advanced PLMs emerge.

6 Conclusion

We propose PNP-VQA, a framework with zero additional training for zero-shot VQA by conjoining off-the-shelf pretrained models. PNP-VQA leverages an image-question matching module to determine image patches relevant to the current question. An image captioning module then generates question-guided captions, which are processed by a question answering module to produce an answer. PNP-VQA achieves state-of-the-arts performance on multiple VQA benchmarks. We hope that our work will bring inspiration for further research in flexible, modular AI systems for solving vision-language tasks.

7 Limitations

Like two sides of the same coin, the strengths and weaknesses of PNP-VQA both result from the zero-training modular system design. PNP-VQA enjoys the power of pretrained models but also inherits the bias from these models. It enjoys the efficiency of zero training, but introduces additional inference cost due to the multi-step process. Nevertheless, we believe that the strengths of PNP-VQA outweigh its limitations, and welcome further investigations to help debias pretrained models and improve inference speed.

8 Acknowledgments

Anthony Meng Huat Tiong is supported by Salesforce and Singapore Economic Development Board under the Industrial Postgraduate Programme. Boyang Li is supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

References

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. 2022. Flamingo: a visual language model for few-shot learning. *arXiv Preprint 2204.14198*.

Jacob Andreas, Dan Klein, and Sergey Levine. 2018. Learning with latent language. 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 2166–2179, New Orleans, Louisiana. Association for Computational Linguistics.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. WeaQA: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, Online. Association for Computational Linguistics.

Maxwell A. Bertolero, B. T. Thomas Yeo, and Mark D'Esposito. 2015. The modular and integrative functional architecture of the human brain. *Proceedings of the National Academy of Sciences*, 112(49):E6798–E6807.

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

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An

open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *arXiv Preprint 2108.07258*.

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. *arXiv Preprint 2201.06028*.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. Language gans falling short. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. In *North American Chapter of the Association for Computational Linguistics*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18030–18040.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. 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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. Magma – multimodal augmentation of generative models through adapter-based finetuning. *arXiv Preprint 2112.05253*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. 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. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Jerry Alan Fodor. 1983. *The Modularity of Mind*. MIT Press.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. 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. Association for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. UnifiedQA-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2022a. Lavis: A library for language-vision intelligence.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

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. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *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.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: a visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Sara J Shettleworth. 2012. Modularity, comparative cognition and human uniqueness. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1603):2794–2802.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5227–5237.

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

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wai Keen Vong and Brenden M. Lake. 2022. Few-shot image classification by generating natural language rules. In *ACL Workshop on Learning with Natural Language Supervision*.

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

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations, ICLR 2022*. OpenReview.net.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv Preprint 2109.01652*.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022a. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022b. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

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

A Visualization

In the appendix, we show visualizations of GradCAM heatmaps and the generated captions for VQAv2, OK-VQA, and GQA in following pages.

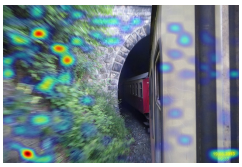
B Hyperparameter sensitivity

We study how VQAv2 validation accuracy varies with different cross-attention layer used for GradCAM and number of image patches sampled for question-guided caption generation. Figure 9(a) shows no clear relationship between VQA accuracy and the cross-attention layer used for GradCAM. The maximum difference in VQA accuracy across different cross-attention layers is 3%. Figure 9(b) shows that VQA accuracy has a negative correlation with the number of sampled image patches. As K' increases, the sampled patches become less relevant to the questions, and question-guided patch sampling becomes akin to using all patches.

Q: is this picture blurry? A: yes

Generic captions:
1. a view of a train going through a tunnel
2. a train tunnel next to a tree filled forest

Prediction: no



Question-guided captions:
1. the view of a blurry image of trees and bushes
2. motion picture of a train window with blurred photo

Prediction: yes

Q: what color is the shoe? A: red

Generic captions:
1. a dog that is on the floor with shoes next to it
2. a brown and gray dog sleeping under a table

Prediction: black

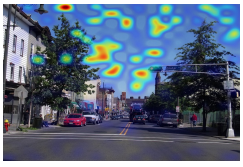
Question-guided captions:
1. a pair of red trainers and a pair of red shoes are shown
2. a red sneakers and red boot and a pair of red shoes are pictured next to

Prediction: red

Q: are clouds visible? A: no

Generic captions:
1. a small town with cars parked along a one way street
2. cars and a vehicle at a red light on a corner

Prediction: yes



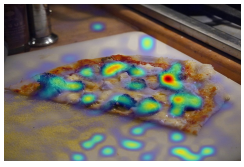
Question-guided captions:
1. cars stopped at an intersection on a clear day
2. the clear blue sky above is blue in color and is a clear sky

Prediction: no

Q: what is the green stuff on top? A: broccoli

Generic captions:
1. a slice of pizza on a cutting board with cheese and broccoli
2. a vegetarian pizza with parsley on a marble dinner plate

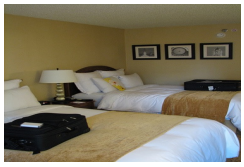
Prediction: parsley



Question-guided captions:
1. slices of broccoli and cheese on a pizza pie
2. a picture of some slices of broccoli, broccoli and pizza

Prediction: broccoli

Q: is there any art hanging on the walls? A: yes



Generic captions:
1. two beds in a suite with luggage in a bag on top of them
2. two large beds sitting in a room with suitcases

Prediction: no

Question-guided captions:
1. three pictures in a frame above two beds
2. a hotel room with 2 double beds and pictures on the wall

Prediction: yes

Q: what is the name of the theater? A: grand

Generic captions:
1. a very tall tower with a little clock on it
2. there is an old clock tower at this town

Prediction: the palace

Question-guided captions:
1. a white grand theatre, on a bright day
2. the grand store, grand in grand, is seen

Prediction: grand

Q: what number is the horse? A: 6

Generic captions:
1. a man walking a black horse on a track
2. a horse with a number 9 in a race track

Prediction: 9

Question-guided captions:
1. a jockey is on his horse and numbers are on the number 6
2. a jockey is taking a pony with the name number six eight

Prediction: 6

Q: what color is the line on the tennis court? A: white

Generic captions:
1. a tennis player is playing tennis on a red court
2. a tennis player getting ready to serve the ball

Prediction: red

Question-guided captions:
1. a tennis court with a red clay tennis court and white line
2. a woman on a clay tennis tennis court preparing to strike a ball

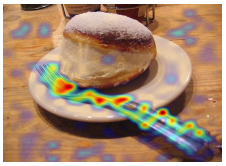
Prediction: white

Figure 6: Examples from **VQAv2**. We show generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues in captions. Words in red indicate wrong answer predictions.

Q: what utensil is this?
A: fork

Generic captions:
1. a spoon and fork are sitting on a white plate on a wooden table
2. a round cake with cream on it on a plate

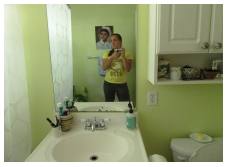
Prediction: a spoon



Question-guided captions:
1. a fork, silverware, fork and a spoon are shown
2. utensil on the plate which seems to have a fork and the fork

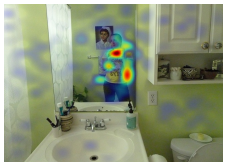
Prediction: fork

Q: what is the popular name for the type of photo this lady is taking? A: selfie



Generic captions:
1. a smiling teen girl taking a picture in a mirror
2. a person standing in a small bathroom taking a photo

Prediction: self-portrait



Question-guided captions:
1. a woman is taking a selfie and taking a selfie
2. a woman is taking a picture in a mirror and taking a picture

Prediction: selfie

Q: what holiday is being celebrated?
A: christmas

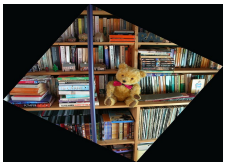
Generic captions:
1. a man at a table with a box and donuts
2. a man with a hat showing a box with a dozen donuts in it

Prediction: donut holiday

Question-guided captions:
1. a man is wearing a santa hat eating a kriskin
2. a person with a red hat eating a donut

Prediction: christmas

Q: what shape is cut out here?
A: diamond



Generic captions:
1. a small stuffed bear sits on a bookshelf with shelves full of books
2. a teddy bear that is sitting in front of a bookcase

Prediction: bear

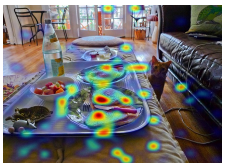
Question-guided captions:
1. a photograph which looks like a diamond pattern
2. square image surrounded by book - diamond diamond diamond

Prediction: diamond

Q: what style breakfast is this?
A: continental

Generic captions:
1. a silver metal tray that has breakfast on top of it
2. a metal tray holding a couple plates of food

Prediction: restaurant



Question-guided captions:
1. pancakes pancakes are shown at place on a set at a dinner table
2. plates of delicious breakfast on a tray of a table

Prediction: continental

Q: what brand of soda is on the bottle?
A: coca cola

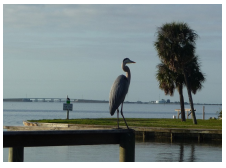
Generic captions:
1. a man in a helmet is standing next to a table
2. a person is standing in a room with a helmet on

Prediction: tequila

Question-guided captions:
1. a man is wearing a helmet with a coca cola cola cola soda soda coke glass
2. a man with sunglasses waves at the camera

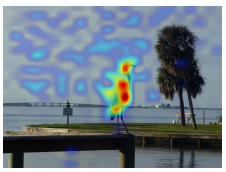
Prediction: coca cola

Q: what species of bird is this?
A: heron



Generic captions:
1. a gray bird is standing on a bench looking out the water with an island in
2. a grey bird in a body of water

Prediction: gray bird



Question-guided captions:
1. a big grey heron is standing up in the sun
2. a heron bird standing in water with heron bird standing next to him

Prediction: heron

Q: what brand of airplane is shown?
A: cessna

Generic captions:
1. a plane sits on a runway at the airport
2. a small white airplane on the runway with trees behind

Prediction: a small white airplane

Question-guided captions:
1. airplane small airplane blue airplane cessna small a airplanes this propeller cessna airplane white private fuselage
2. the small airplane is parked on the runway

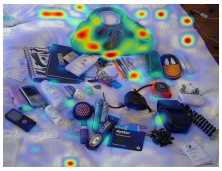
Prediction: cessna

Figure 7: Examples from **OK-VQA**. We show generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues in captions. Words in red indicate wrong answer predictions.

Q: which color is the bag which is lying on top of the bed? A: green

Generic captions:
1. lots of assorted items are on the beds
2. womens personal and personal items laid out for a picnic

Prediction: black



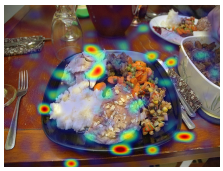
Question-guided captions:
1. a green bag of essentials laid out on white background in the middle of shot
2. a purse sitting on the bed next to several large items

Prediction: green

Q: How is the meat on the plate that is sitting atop the table called? A: turkey

Generic captions:
1. a meal on a plate with meat, vegetables and rice
2. various pieces of food are placed on a plate

Prediction: a meal



Question-guided captions:
1. is thanksgiving turkey dinner roast meal chicken dinner dinner turkey pork dining meal roast chicken bone
2. plate of food is laying on a wood table

Prediction: turkey

Q: what is the crate made of? A: plastic

Generic captions:
1. two men making a blend in a glass bottle
2. two men sitting on a beach next to a plastic container of liquid

Prediction: wood

Question-guided captions:
1. a large bald man looking at a camera and an open plastic crate near by a
2. a juice pours from a beverage in a crate next to yellow crates

Prediction: plastic

Q: is the wide street made of cobblestone? A: yes

Generic captions:
1. a double - decker bus in front of a large white house
2. a bus is filled with a striped design on the street

Prediction: no

Question-guided captions:
1. a double decker bus is riding down the cobblestone brickstone road
2. people on a brick paved cobble stone road

Prediction: yes

Q: what is located on top of the pancake? A: butter

Generic captions:
1. a breakfast plate of pancakes, eggs, and bacon on it
2. pancakes and scrambled eggs on a large white plate with bacon

Prediction: bacon

Question-guided captions:
1. a plate of pancakes with a butter on top
2. breakfast plate with pancakes, syrup, bananas, and butter

Prediction: butter

Q: what kind of watercraft is colorful? A: boats

Generic captions:
1. a man is holding a surfboard and walking across a sandy lot
2. a man holding a baseball bat on top of a sandy beach

Prediction: surfboard

Question-guided captions:
1. a boy standing with blue and green boats at the beach
2. a man standing next to a bunch of boats

Prediction: boats

Q: what is the fence made of? A: wood

Generic captions:
1. a group of well dressed motorcycles lined up near a fence
2. a group of people outside with motorcycles

Prediction: chain link

Question-guided captions:
1. a wood fence is near the fence
2. there is a group of motorbikes parked in front of a fence

Prediction: wood

Q: what kind of appliance is to the left of the oven? A: refrigerator

Generic captions:
1. the kitchen has an oven, dishwasher, and kitchen cabinets
2. a kitchen with a refrigerator, range and oven

Prediction: dishwasher

Question-guided captions:
1. this is a picture of a kitchen with a refrigerator
2. there is some fridges stainless steel refrigerator freezer

Prediction: refrigerator

Figure 8: Examples from **GQA**. We show generic captions (from all patches) based on the original image and question-guided captions (from the sampled patches) based on the GradCAM heatmaps. For illustrative purposes, we highlight words in green to indicate correct answer predictions and the cues in captions. Words in red indicate wrong answer predictions.

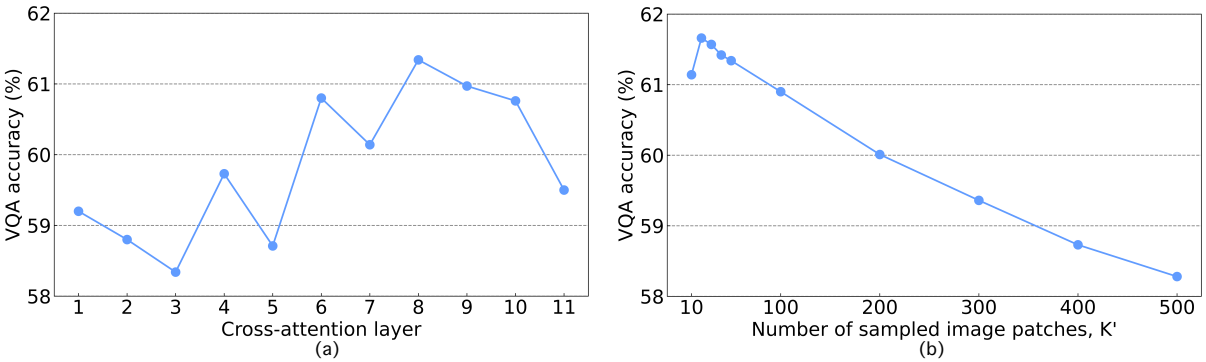


Figure 9: VQAv2 validation set accuracy using (a) different cross-attention layer on which GradCAM is computed using $K' = 50$. (b) different number of image patches sampled for caption generation using GradCAM computed at 8^{th} cross-attention layer.