Self-Bootstrapped Visual-Language Model for Knowledge Selection and Question Answering

Anonymous ACL submission

Abstract

While large pre-trained visual-language models have shown promising results on traditional visual question answering benchmarks, it is still challenging for them to answer complex VQA problems which requires diverse world knowledge. Motivated by the research of retrieval-augmented generation in the field of natural language processing, we use Dense Passage Retrieval (DPR) to retrieve related knowledge to help the model answer questions. However, DPR conduct retrieving in natural language space, which may not ensure comprehensive acquisition of image information. Thus, the retrieved knowledge is not truly conducive to helping answer the question, affecting the performance of the overall system. To address this issue, we propose a novel framework that leverages the visual-language model to select the key knowledge retrieved by DPR and answer questions. The framework consists of two modules: Selector and Answerer, where both are initialized by the MLLM and parameter-efficiently finetuned by self-bootstrapping: find key knowledge in the retrieved knowledge documents using the Selector, and then use them to finetune the Answerer to predict answers; obtain the pseudo-labels of key knowledge documents based on the predictions of the Answerer and weak supervision labels, and then finetune the Selector to select key knowledge; repeat. Our framework significantly enhances the performance of the baseline on the challenging open-domain Knowledge-based VQA benchmark, OK-VQA, achieving a state-of-the-art accuracy of 62.83%.

1 Introduction

Recently, there has been an impressive advancement in large visual-language models (LVLM) (Li et al., 2023; Alayrac et al., 2022; Liu et al., 2023; Dai et al., 2023). They usually use a mapping network to inject visual features into the semantic space of the large language model (Brown et al., 2020; Zhang et al., 2022; Touvron et al., 2023; vic, 2023; Touvron et al., 2023) and demonstrate strong capabilities on multimodal perception and reasoning. Thus, they achieve significant progress in conventional visual question answering benchmarks (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019) which primarily focus on addressing straightforward questions that only necessitate visual perception and recognition. However, it is still challenging for the LVLMs to answer visual questions which require broader world knowledge and common sense (Wang et al., 2017; Marino et al., 2019; Schwenk et al., 2022).

Motivated by the research of retrieval-augmented generation (Karpukhin et al., 2020a) in the field of natural language processing, we use Dense Passage Retrieval (DPR) to retrieve related world knowledge to help the model answer questions. However, when using DPR, we need to transform the image into texts to retrieve the related knowledge, which leads to the underutilization of visual information. Thus, the retrieved knowledge may be unfaithful and affects the model performance. To address the issue, we consider the LVLM as the knowledge selector to find helpful knowledge from candidate retrieved knowledge by DPR. Then the selected knowledge is fed into the LVLM to predict the answer.

In this paper, we introduce a novel framework where we adopt the visual-language model to perform knowledge selection and question answering. Our framework comprises two modules: a Selector and an Answerer. We train two modules by repeating the following process: the Selector first identifies important knowledge from the candidate knowledge documents retrieved by the pre-trained retriever; then, the Answerer takes the key knowledge documents as the input knowledge and is finetuned to generate the answer; next, we generate pseudo-labels of key knowledge documents according to the Answerer's predictions and weak supervision labels; finally, we refine the Selector to assess the relevance of retrieved knowledge documents in answering the question. This strategy of self-bootstrapping enhances the ability of knowledge selection and answer generation consistently, enabling the model to accurately respond to knowledge-intensive questions.

We conduct extensive experiments on the open-domain knowledge-based VQA benchmark (OK-VQA (Marino et al., 2019)) to validate the effectiveness of the proposed framework, where our method largely outperforms the baseline and achieves the state-of-the-art performance of 62.83%, only fine-tuning 0.16% parameters with LoRA (Hu et al., 2022a). We also conduct comprehensive ablations to validate the impact of different components of the proposed framework, including the Effect of Selector and Answerer, cycle training of the framework, varying the number of key knowledge documents, the impact of vision information, and so on.

Our contributions are summarized as follows:

- We introduce a novel framework that leverages the large visual-language model to select key knowledge and use them to answer questions, respectively.
- We propose a new self-bootstrap learning method to train the Selector and Answerer, where the Selector chooses key knowledge documents for the Answerer and the Answerer provides pseudo-labels for the Selector.
- We achieve a state-of-the-art performance of 62.83% on the OK-VQA dataset, surpassing the previous state-of-the-art method. Notably, this improvement is achieved by fine-tuning only 0.16% of parameters using LoRA.

2 Related work

Large Visual-Language Models. Recently, large visual-language models (Li et al., 2023; Alayrac et al., 2022; Liu et al., 2023; Dai et al., 2023) have demonstrated remarkable visual-language understanding and reasoning capabilities, owing to the advancement of larger language models (Brown et al., 2020; Zhang et al., 2022; Touvron et al., 2023; vic, 2023; Touvron et al., 2023). These methods typically consist of a frozen visual encoder (Radford et al., 2021), a visual-language connector (Li et al., 2023), and a large language model (Chung et al., 2022; Zhang et al., 2022; vic, 2023). The models are firstly pre-trained on large-scale visual-text datasets to align visual features to the language embedding space. After pretraining, the large language model can understand the visual details. Then, the model is finetuned to adapt to various visual-language tasks. In this study, we adopt BLIP2, one of the widely used models, as our backbone for bootstrapping knowledge selection and question answering with it.

Knowledge-based VQA. Conventional VQA benchmarks (Goyal et al., 2017; Hudson and Manning, 2019) primarily focus on basic visual perception and reasoning tasks and numerous studies have achieved promising results on these benchmarks (Anderson et al., 2017; Zhang et al., 2021; Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2022; Wang et al., 2022). Different from them, the knowledge-based VQA task (Wang et al., 2017; Marino et al., 2019; Schwenk et al., 2022) requires models to incorporate diverse world knowledge to respond to questions about visual content, which is more challenging. Recent studies (Gardères et al., 2020; Wu et al., 2022; Lin and Byrne, 2022; Gui et al., 2021) have explored various open-domain world knowledge sources, such as ConceptNet (Speer et al., 2017), Wikipedia (Vrandečić and Krötzsch, 2014), Google Search Corpus (Luo et al., 2021). They retrieve the relevant knowledge documents from the knowledge bases and integrate them into the answering model to generate predictions. Except for using explicit knowledge, some methods also take GPT-3 (Brown et al., 2020) as an implicit knowledge producer. They either prompt GPT-3 with in-context examples to predict answers directly (Yang et al., 2022; Hu et al., 2022b; Shao et al., 2023), or use GPT-3 to generate answer candidates with evidence serving as textual implicit knowledge bases (Gui et al., 2021; Lin et al., 2022), leading to significant performance improvements. Different from these approaches, we employ a large visual-language model to select key retrieved knowledge and reason on the knowledge to answer questions.

3 Method

In this section, we first introduce the preliminaries of Knowledge Retrieval and LVLM, which are the foundation of our framework. Then, we present the design of the Selector and Answerer for knowledge selection and question answering on knowledge respectively. Finally, we illustrate the self-bootstrap training method of two designed modules.

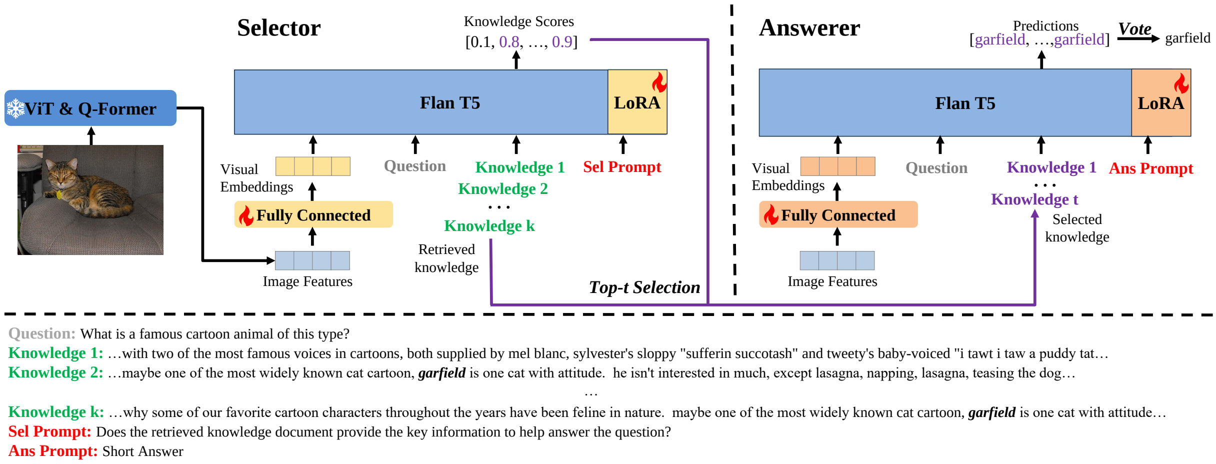

Figure 1: Our framework consists of two modules: a Selector and an Answerer. Selector (left) selects the top-T knowledge documents for the Answerer (right), and the Answerer focuses on important knowledge information to predict answers. Both modules utilize the same frozen visual module to extract image features. We train the fully connected (FC) layer and fine-tune the language model using LoRA, which amounts to only 0.16% of the total parameters. For detailed training procedures of the two modules, refer to Alg. 1. The original knowledge is retrieved using DPR, and for brevity, we omit the retrieval process here (details can be found in Section 3.1).

3.1 Preliminaries

Knowledge Retrieval. We adopt the Dense Passage Retrieval (DPR) (Karpukhin et al., 2020b) to retrieve the knowledge documents. We transform the image into raw texts composed of captions, objects, attributes, and OCR (Optical Character Recognition). Then we compute the similarity scores between the query and knowledge documents $sim(q_i, D_j) = \mathbf{q}_i^T \cdot \mathbf{d}_j$ and exploit FAISS (Johnson et al., 2019) to index Top-k related knowledge documents $\mathcal{P}_i = \{P_{i,1}, P_{i,2}, ..., P_{i,k}\}$ for i-th query.

Large Visual-Language Model. In our work, both knowledge selection and question-answering modules adopt BLIP-2 (Li et al., 2023) as the backbone. The architecture of BLIP-2 comprises a frozen image encoder (Dosovitskiy et al., 2020; Fang et al., 2023), a Q-Former (Li et al., 2023), and a pre-trained language model (Chung et al., 2022). Given an image I_i, the frozen image encoder outputs a set of visual features $\{\mathbf{h}_{i,1}, \mathbf{h}_{i,2}, ..., \mathbf{h}_{i,m}\}$. Q-Former takes extracted visual features as input, and outputs language-aligned visual features $\{\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, ..., \mathbf{v}_{i,l}\}$. These visual features are concatenated with the textual word embeddings, which are fed into the language model for generation. Through pre-training on large-scale image-caption datasets, Q-Former can effectively project visual features into the feature space of the Language Large Model (LLM). We freeze the visual encoder and Q-former during training. We train the fully connected layer and use LoRA (Hu et al., 2022a) to finetune the LLM (only finetune 0.16% of total parameters).

3.2 Selector and Answerer

Selector. After obtaining the Top-k knowledge documents using DPR for the i-th sample, we aim to choose t most important knowledge documents from the retrieved documents. where t is smaller than k. As shown in Fig. 1, we firstly use the frozen image encoder and Q-former to extract the image features $\mathbf{V}_i$, where these features are extracted once and then used by the Selector and the Answerer. Then image features $\mathbf{V}_i$ are fed into the independent fully-connected layer to obtain the visual embeddings $\mathbf{E}_i^v$. We concatenate the question, a retrieved knowledge document, and the Selection prompt "Does the retrieved knowledge document provide the key information to help answer the question?" into one sentence S. Next, visual embeddings $\mathbf{E}_i^v$ and the text are concatenated and fed into the LLM (Flan-T5 (Chung et al., 2022) is adopted in our work). Last, we use the probability of generating the word 'yes' as the score of each retrieved knowledge document $P_{i,j}$, denoted as $s_{i,j} = \boldsymbol{LLM}(concat(\mathbf{E}_i^v, S_i))$, and we select top-t documents $\hat{\mathcal{P}}_i = \{\hat{P}_{i,1}, \hat{P}_{i,2}, ..., \hat{P}_{i,t}\}$ based on the scores. The Selector can be conceptualized as follows:

$$\hat{\mathcal{P}}_i == Selector(I_i, Q_i, \mathcal{P}_i), |\hat{\mathcal{P}}_i| = t \quad (1)$$

Answerer. After obtaining the selected knowledge

Algorithm 1 Pipeline of cycle training

Input:
KB-VQA dataset $\mathcal{D} = \{I_i, Q_i, \mathcal{A}_i | i = 1, 2, \ldots, N\}$;
Retrieved knowledge documents $\mathcal{P}_i = \{P_i^1, P_i^2, \ldots, P_i^k\}$; I_i, Q_i, $\mathcal{P}_i$, and $\mathcal{A}_i$ denote image, question, document set, and answer set of i-th sample
Output: Knowledge selection model $Selector$; Question answering model $Answerer$
for sample in $\mathcal{D}$ **do**
  Stage 1:
  1: Using $Selector$ to select top-t documents $\hat{\mathcal{P}}_i$ from the retrieved knowledge documents $\mathcal{P}_i$ as Eq. 1
  2: Finetuning $Answerer$ on $\{I_i, Q_i, \hat{\mathcal{P}}_i, \mathcal{A}_i\}$ supervised by the ground-truth answer as Eq. 3.
  Stage 2:
  1: Using $Answerer$ to predict answers for retrieved knowledge documents $\mathcal{P}_i$ as Eq. 2
  2: Generating to pseudo labels $\{y_{i,j}\}$ for retrieved knowledge documents $\mathcal{P}_i$ as Eq. 4
  3: Finetuning $Selector$ on $\{I_i, Q_i, \mathcal{P}_i, \{y_{i,j}\}\}$ supervised by the pseudo label as Eq. 5.
end for

documents, we aim to reason on the knowledge to answer questions. As shown in Fig. 1, we process the same image features to obtain the different visual embeddings $\mathbf{E}_i^v$ via the fully-connected layer of the Answerer. Next, we concatenate the question and the knowledge into one sentence S' using the template "Question: {} Knowledge: {} Answer: ". We concatenate the visual embeddings and the text, which are fed into the LLM with different LoRA parameters to get the answer. The model outputs corresponding answers based on different documents. The Answerer can be conceptualized as follows:

$$a_i = Answerer(I_i, Q_i, \hat{\mathcal{P}}_i) \tag{2}$$

Then the final answer is based on the majority vote. We also tried different knowledge reasoning methods, such as concatenating (the results can be seen in the ablation study).

3.3 Self-Bootstrap Learning

To enable the Selector and Answerer to select key knowledge and answer questions, we bootstrap them with each other in a style of cycle training. We repeat the following process for the given i-th sample $\{I_i, Q_i, \mathcal{P}_i, \mathcal{A}_i\}$ of the training dataset:

Answerer Training. We use Eq. 1 to get the selected knowledge documents $\hat{\mathcal{P}}_i$. The image I_i is fed into the frozen ViT and Q-former to obtain the image features $\mathbf{V}_i$. We use the trainable FC_{ans} layer to output the visual embeddings $\mathbf{E}^v_{ans,i}$. We concatenate the visual embedding, the question Q_i and each selected knowledge document $\hat{P}_{i,j}$ to construct t triplets for the sample, where $j = 1, 2, \ldots, t$. Then we finetune the Answerer with LoRA under the supervision of the ground truth answer $\mathcal{A}_i$:

$$\begin{aligned} \mathbf{E}^v_{ans,i} &= FC_{ans}(\mathbf{V}_i), \\ L_{ans} &= -\sum_{j=1}^{t} \log LLM_{ans}(a_i^* | \mathbf{E}^v_{ans,i}, Q_i, \hat{P}_i^j), \end{aligned} \tag{3}$$

where a_i^* is the most frequent answer in the human-annotated answer set $\mathcal{A}_i$.

Selector Training. We first use Eq. 2 to predict answers based on each retrieved knowledge document $P_{i,j}$. Then we assign pseudo labels to the retrieved documents according to model predictions and weak supervision labels (Luo et al., 2021; Lin and Byrne, 2022; Lin et al., 2023). We use "yes" and "no" as pseudo labels, where label a document as positive knowledge if Answerer can output the correct answer using that document and the document contains any of the answers in $\mathcal{A}_i$.

$$y_{i,j} = \begin{cases} \text{yes}, & \textit{if } \ a_i = a_i^* \wedge \\ & \quad P_{i,j} \ \text{contains an answer in} \ \mathcal{A}_i \\ \text{no}, & \textit{else} \end{cases} \tag{4}$$

After obtaining the pseudo label of each retrieved knowledge document, we use the trainable FC_{sel} layer to output the visual embeddings $\mathbf{E}^v_{sel,i}$. we concatenate the visual embedding, the question Q_i and each retrieved knowledge document $P_{i,j}$ to construct k triplets for the sample, where $j = 1, 2, \ldots, k$. Then we finetune the Selector with LoRA under the supervision of pseudo labels:

$$\begin{aligned} \mathbf{E}^v_{sel,i} &= FC_{sel}(\mathbf{V}_i), \\ L_{sel} &= -\sum_{j=1}^{k} \log LLM_{sel}(y_{i,j} | \mathbf{E}^v_{sel,i}, Q_i, P_i^j) \end{aligned} \tag{5}$$

Table 1: **Performance comparison with state-of-the-art (SOTA) methods on the OK-VQA dataset.** Knowledge Sources: **C**onceptNet (C); **W**ikipedia (W); **G**oogle **S**earch (GS); **G**oogle **I**mages (GI). The best result in the table is bolded. The results show that our method achieves the state-of-the-art performance.

Models	Large Models	K_{train}	K_{test}	Knowledge Resource	Accuracy (%)
BAN+AN (Marino et al., 2019)	-	-	-	W	25.6
ConceptBERT (Gardères et al., 2020)	-	-	-	C	33.7
KRISP (Marino et al., 2021)	-	-	-	C+W	38.4
Visual Retriever-Reader (Luo et al., 2021)	-	100	100	GS	39.2
MAVEx (Wu et al., 2022)	-	-	-	W+C + GI	39.4
PICa (Yang et al., 2022)	GPT-3 (175B)	-	-	GPT-3	48.0
TRiG(Ensemble) (Gao et al., 2022)	T5-large (770M)	100	100	W	50.5
KAT(Single) (Gui et al., 2021)	T5-large (770M)	40	40	W + GPT-3	53.1
KAT(Ensemble) (Gui et al., 2021)	T5-large (770M)	40	40	W + GPT-3	54.4
RA-VQA (Lin and Byrne, 2022)	T5-large (770M)	5	50	GS	54.5
REVIVE(Single) (Lin et al., 2022)	T5-large (770M)	40	40	W+GPT-3	56.6
REVIVE(Ensemble) (Lin et al., 2022)	T5-large (770M)	40	40	W+GPT-3	58.0
PromptCap (Hu et al., 2022b)	GPT-3 (175B)	-	-	GPT-3	60.4
Prophet (Shao et al., 2023)	GPT-3 (175B)	-	-	GPT-3+MCAN	61.1
FillingGap (Wang et al., 2023)	GPT-3 (175B)	-	-	GPT-3	61.3
SimpleBaseline (Xenos et al., 2023)	LLaMA 2 (13B)	-	-	LLaMA 2	61.2
Cola-FT (Chen et al., 2024)	FLAN-T5(11B)	-	-	BLIP+OFA	62.4
Flamingo (Alayrac et al., 2022)	Flamingo (80B)	-	-	Pretrain	57.8
InstructBLIP (Dai et al., 2023)	InstructBLIP Vicuna (7B)	-	-	Pretrain	62.1
Qwen-VL (Bai et al., 2023)	Qwen-VL(Qwen-7B)	-	-	Pretrain	58.6
MM-Reasoner (Khademi et al., 2023)	Flamingo (80B)	-	-	GPT-4	60.8
BLIP2 (fine-tuned) (Li et al., 2023)	BLIP2 T5-XL (3B)	-	-	Pretrain	55.44
RA-VQA-v2 (Lin et al., 2023)	BLIP2 T5-XL (3B)	5	5	GS	62.1
PreFLMR (Lin et al., 2024)	BLIP2 T5-XL (3B)	5	5	GS	61.88
Ours	BLIP2 T5-XL (3B)	5	5	GS	**62.83**

We provide the overall training pipeline in Alg. 1. Through continuous iteration, the Selector will provide more crucial knowledge for the Answerer to accurately respond to questions. Meanwhile, the improvement in the Answerer's reasoning ability will also result in more precise pseudo-labeling, further enhancing the Selector's discriminative power. During the inference stage, we utilize the Selector to choose key knowledge, and then instruct the Answerer to respond to questions based on this knowledge.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct extensive experiments on OK-VQA (Marino et al., 2019) to evaluate the effectiveness of our method. OK-VQA is a challenging open-domain knowledge-based VQA dataset that requires models to leverage various external knowledge sources to answer questions. The dataset contains 14,055 questions and 14,031 images, whereas the training set and testing set have 9k and 5k image-question pairs, respectively. Due to no knowledge base being provided for OK-VQA, we need to choose the proper knowledge base for the dataset. In this paper, we adopt Google Search Corpus (Luo et al., 2021) as the knowledge base which is collected in the websites using the Google Search API.

Evaluation Metric. We use the standard VQA metric (Antol et al., 2015) to evaluate the performance of the model. Given the prediction of the question a and the groudtruth answer set $\mathcal{A}$, the VQA accuracy is calculated as:

$$Accuracy(a, \mathcal{A}) = \min(\frac{\#A(a)}{3}, 1), \qquad (6)$$

where the groudtruth answer set $\mathcal{A}$ is annotated by different humans, $\#A(a)$ denotes the occurrence of a in $\mathcal{A}$.

Implementation Details. In our experiment, we adopt BLIP2 T5-XL (3B) (Li et al., 2023) to initialize the Selector and Retriever. We freeze the image encoder and Q-former, with both the Selector and Retriever sharing the same visual module. We finetune the fully connected layer and use LoRA (Hu et al., 2022a) to train the LLM. We use the default huggingface-PEFT setting: r=8, lora_alpha=32, lora_dropout=0.1. We use Adam as the optimizer and set the batch size to 8. We use the warm-up strategy which trains the model with an initial learning rate of 1e-4 and warm-up factor of 0.05 for 1000 steps and then utilizes a cosine annealing learning strategy with an initial learning rate of 1e-4 and a final learning rate of 0 after 10 epochs. We use top-30 knowledge documents retrieved by a pre-trained DPR (Karpukhin et al., 2020b) as candidates for Selector and use the selected top-5 documents from the 30 documents for the Answerer to train and infer, denoted as $K_{candidate} = 30, K_{train} = 5, K_{test} = 5$. We use 2 Nvidia A800 GPUs (80G) for all experiments. And our codes will be released upon paper acceptance.

4.2 Comparison with State-of-the-art Methods

As shown in Tab. 1, we can see early models (BAN+AN (Marino et al., 2019), ConceptBERT (Gardères et al., 2020), KRISP (Marino et al., 2021), Visual Retriever-Reader (Luo et al., 2021), and MAVEx (Wu et al., 2022)) have a weak performance, achieving a VQA accuracy from 25.6% to 39.4%. Recently, by introducing larger models (T5-large, GPT-3, LLaMA, Vicuna) and diverse knowledge resources (ConceptNet, Wikipedia, Google Web Search and Google Images), the performance has a significant performance improvement, achieving a VQA accuracy of 62.4%. Our method aims to augment the reasoning ability to answer knowledge-intensive questions of the large visual-language model. When directly finetuning BLIP2 T5-XL on OKVQA, the model has a low performance of 55.44%. By introducing external knowledge, the performance has a significant performance improvement. Different from RA-VQA-v2 (Lin et al., 2023) and PreFLMR (Lin et al., 2024), we do not train a multimodal retriever from scratch which requires expensive annotations and high computational costs. We directly leverage the large visual-language model to select key knowledge from the retrieved knowledge by DPR like the process of re-ranking. With the same knowledge resources (*i.e.,* Google Search), our method can achieves 62.83% accuracy, outperforming other state-of-the-art models. It is worth noting that we do not use GPT-3 and we only train the 0.16% parameters of the model. These results demonstrate the effectiveness of the proposed approach.

4.3 Ablation Study

We conduct the ablation studies to evaluate different components of our framework on OK-VQA.

Table 2: Ablation study on the Selector. We select 5 knowledge documents from top-30 knowledge candidates retrieved by DPR. 'DPR Score' refers to selecting top-5 knowledge based on similarity scores. 'Random Selection' means randomly selecting 5 knowledge documents from 30 candidate knowledge documents. 'Selector' denotes choosing 5 key knowledge documents by the Selector.

K_{train}	K_{test}	Knowledge Selection	Accuracy (%)
5	1	Random Selection	50.45
5	1	DPR Score	58.80
5	1	Selector	61.62
5	5	Random Selection	55.05
5	5	DPR Score	60.69
5	5	Selector	62.83

Effect of Selector. We conduct the ablation study to evaluate the effectiveness of Selector in our method. We show the results in Tab. 2. From the results, we can observe: Our framework, leveraging key knowledge documents selected by the Selector, consistently outperforms the Answerer when using the same number of documents retrieved by DPR. We improve the performance by 2.14% and 1.88% with 1 and 5 test knowledge documents, compared to DPR-based retrieval. When using the randomly selected documents, the model performs worst. These results demonstrate that top-ranked knowledge documents based on DPR scores are not optimal for question answering and our key knowledge selection module can identify relevant documents for accurate question answering, ensuring the coherence of knowledge retrieval and question-answering processes.

Effect of Answerer. In Tab. 3, we present a comparison of our Answerer using different knowledge reasoning methods. The results show that the performance using the strategy of voting surpasses that of concatenating under different knowledge selection settings. We argue that directly combining all the knowledge documents into a lengthened document makes it difficult for the Answerer to reason on them, which is easily influenced by noisy information. In contrast, it is easier for the Answerer to reason on each document to predict the answer. Simple voting can choose the best answer.

Effect of Self-Bootstrap Learning. To evaluate the effectiveness of our self-bootstrap learning method, we compare the method with the strategy of independent training of two modules. We finetune the Answerer with the knowledge docu-

Table 3: Ablation study on different knowledge reasoning methods of Answerer. 'Concatenating' denotes that we combine the key knowledge documents into one sentence and feed it into the Answerer to predict the final answer. 'Voting' means that we feed different key knowledge documents into the Answerer to predict different answers and choose the best answer based on majority voting.

Method	Backbone	Accuracy (%)
Concatenating	BLIP2 (fine-tuned)	59.11
Voting	w knowledge from DPR	60.69
Concatenating	Ours	62.06
Voting	w knowledge from Selector	62.83

Table 4: Ablation study on different training methods of our framework.

Methods	Accuracy (%)
Baseline	60.69
Independent training	59.02
Cycle training	62.83

ments retrieved by DPR as the baseline. Independent training means that we train the Selector and the Answerer respectively. Initially, we train the Answerer module utilizing knowledge documents retrieved by DPR. Subsequently, employing the trained Answerer, we generate answers for each retrieved knowledge document, thereby generating pseudo-labels for the retrieved knowledge. We then proceed to train the Selector module supervised by these pseudo-labels. Finally, we conduct finetuning of the Answerer once more, incorporating new knowledge documents selected by the Selector. The results in Tab. 4 show that the model with cycle training outperforms the model with iterative refinement by 3.81%. The VQA score of using independent training is even lower than the baseline. These results demonstrate that our cycle training method can effectively boost the Selector and Answerer each other, which makes the model find key knowledge documents and leverage the knowledge to answer questions.

Effect of different methods of pseudo-labeling. In Tab. 5, we compare the model performance with different methods of pseudo-labeling. When using the model predictions as guidance, the model has a VQA score of 62.31%. When adding the weak supervision as the guidance, the model's VQA score increases from 62.31% to 62.83%. The results demonstrate that using weak supervision labels preserves potentially useful documents, aiding the Answerer in accurately answering questions.

Table 5: Ablation study on different methods of pseudo-labeling.

Model predictions	Weak supervision labels	Accuracy (%)
✓		62.31
✓	✓	62.83

Table 6: Ablation study on different numbers of candidate documents and selected documents.

$K_{candidate}$	K_{train}	K_{test}	Accuracy (%)
5	1	1	57.90
5	1	5	58.32
10	1	1	58.61
10	1	5	59.40
10	5	5	61.86
15	5	5	62.31
30	5	5	62.83
30	5	1	61.62

Effect of key knowledge documents ranges and quantities. In Tab. 6, we evaluate key knowledge document selection using various numbers of candidate documents and selected documents. From the results, we have the following findings: (1) As the number of selected documents increases, the model's performance improves. This indicates that using more documents to train and test contributes to answering questions. (2) Using more documents for training can improve the performance a lot (the 2nd line $v.s.$ the last line). However, using more documents for testing has almost no improvement (the 3rd line $v.s.$ 4th line). (3) When the number of candidate documents increases, the model's performance improves. The result demonstrates that low-ranked documents based on DPR scores may contain useful information for question answering. It is necessary for the model to select key knowledge documents.

Effect of different knowledge documents selection in Answerer fine-tuning. Tab. 7 compares the

Table 7: Ablation study on different documents selection in Answerer fine-tuning.

Knowledge Selection		Accuracy (%)
Training	Inference	
DPR	Selector	62.31
Selector	DPR	60.75
Selector	Selector	62.83

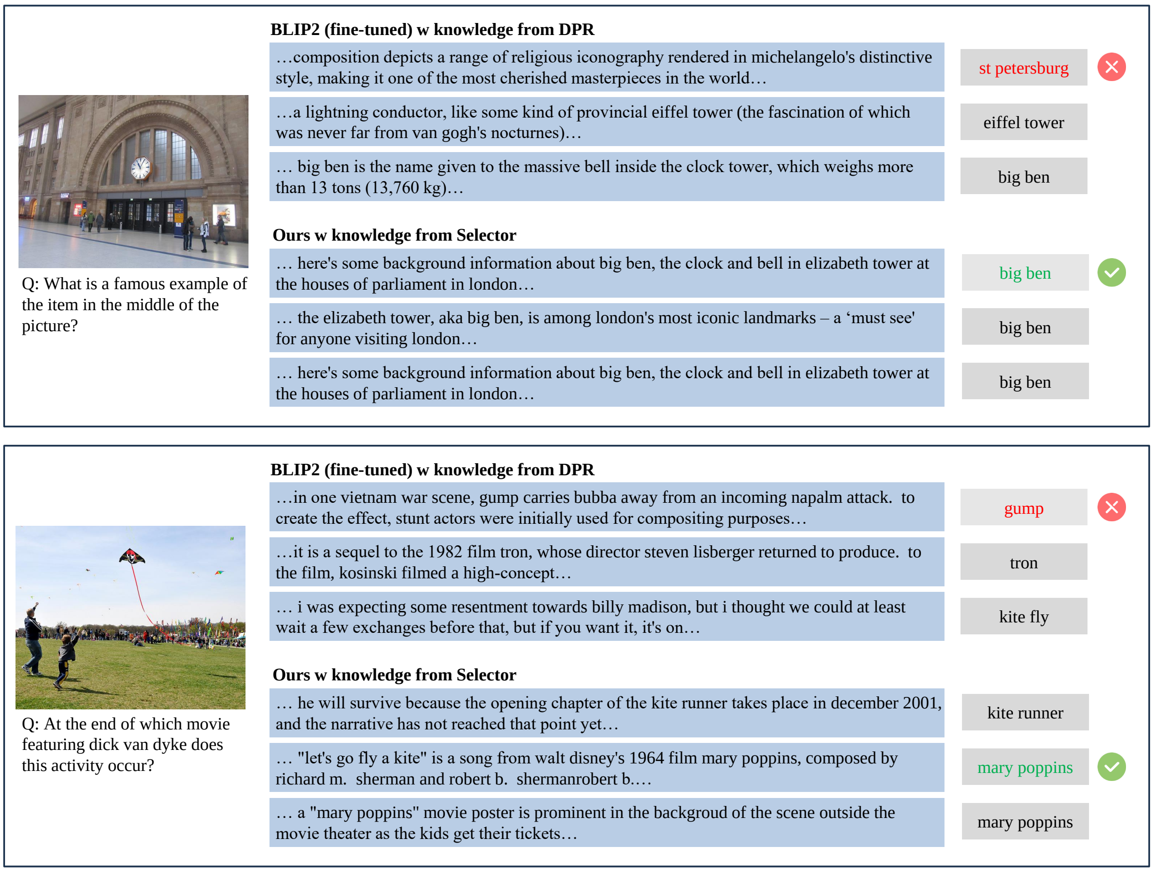

Figure 2: Qualitative results on the test split of OK-VQA. We compared our method with a model that fine-tunes BLIP2 with knowledge ranked by DPR. The middle segment of the graph represents knowledge from various methods used to answer questions. On the right side of the graph, different answers are depicted when using distinct knowledge. Green and red colors indicate whether the selected final answer is correct.

Answerer fine-tuning with different document selection strategies. The results show that our framework performs optimally when utilizing Selector in both Answerer training and inference. This is likely because the Selector provides more informative key knowledge documents and using both Selector ensures the consistency between the training domain and testing domain.

4.4 Qualitative Analysis

In Fig. 2, We present a case study comparing our method with a model that fine-tunes BLIP2 using knowledge ranked by DPR. In the first case, top-ranked knowledge documents from DPR misguide the model, resulting in incorrect predictions. However, our method's Selector chooses key knowledge documents that aid in predicting correct answers. In the second case, each knowledge document from DPR contains irrelevant information, leading to an incorrect final answer. Despite the top-1 document from the Selector resulting in a wrong answer, our method identifies other key knowledge documents for generating correct answers. Through majority voting, the final selected answer is correct. These cases demonstrate our method's ability to extract informative knowledge from retrieved documents to support accurate question answering.

5 Conclusion

In this paper, we propose a novel framework that leverages the large visual-language model to construct two modules: (1) Selector for finding key retrieved knowledge and (2) Answerer for reasoning on the knowledge to predict answers. We design a self-bootstrap learning method to improve their abilities, where the Selector chooses key knowledge documents for the Answerer and the Answerer provides pseudo-labels for the Selector. Compared with state-of-the-art methods, our method achieves better performance on a challenging open-domain knowledge-based VQA benchmark (OK-VQA) and we conduct a comprehensive analysis to evaluate the effectiveness of our method.

6 Limitations

Although our framework can effectively select key knowledge documents for answering question, it is inevitable that the knowledge still contains noise. In some cases, the model itself can answer the question without external knowledge, introducing extra knowledge may affect the performance. In the future, we can explore to dynamically select required knowledge to help itself answer questions.

References

2023. Vicuna. https://github.com/lm-sys/FastChat.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. 2024. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. 2021. Zero-shot visual question answering using knowledge graph. In *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*, pages 146–162. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498.

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*, pages 6904–6913.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. 2022. A unified end-to-end retriever-reader framework for knowledge-based vqa. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2061–2069.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022b. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly supervised concept expansion for general purpose vision models. In *European Conference on Computer Vision*, pages 662–681. Springer.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. MM-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581, Singapore. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *ArXiv*, abs/2309.17133.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. (arXiv:2402.08327).

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *arXiv preprint arXiv:2109.04014*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. 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.

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*, pages 3195–3204.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023. Filling the image information gap for vqa: Prompting large language models to proactively ask questions. *arXiv preprint arXiv:2311.11598*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14871–14877.

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*, volume 36, pages 3081–3089.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Table 8: **Performance comparison with state-of-the-art (SOTA) methods on the FVQA dataset.**

Method	Acc-1
Human	77.99
UnifER (Guo et al., 2022)	55.04
FVQA (Wang et al., 2017)	56.91
ZS-VQA (Chen et al., 2021)	58.27
FVQA(Ensemble) (Wang et al., 2017)	58.76
MM-Reasoner(Ensemble) (Khademi et al., 2023)	61.10
Ours	**63.3**

Table 9: **Performance comparison with state-of-the-art (SOTA) methods on the A-OKVQA dataset.**

Method	Direct Answer	
	val	test
ClipCap (Schwenk et al., 2022)	18.1	15.8
Pythia (Jiang et al., 2018)	25.2	21.9
ViLBERT (Lu et al., 2019)	30.6	25.9
LXMERT (Tan and Bansal, 2019)	30.7	25.9
KRISP (Marino et al., 2021)	33.7	27.1
GPV-2 (Kamath et al., 2022)	48.6	40.7
BLIP-2 T5-XL (Li et al., 2023)	53.2	49.7
PromptCap + GPT-3 (Hu et al., 2022b)	56.3	**59.6**
Ours	**57.2**	56.4

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendix

A.1 Experiments on other datasets.

We also evaluate our method on FVQA (Fang et al., 2023) and A-OKVQA (Schwenk et al., 2022) to demonstrate the effectiveness of our method. FVQA is a VQA dataset that mostly contains questions requiring external knowledge to answer, and provides supporting fact triplets alongside the image-question-answer triplets. A-OKVQA is an augmented successor of OK-VQA, containing 25K image-question pairs that require broader commonsense and world knowledge to answer. Due to A-OKVQA does not provide the knowledge source, we use Wikipedia (Vrandečić and Krötzsch, 2014) as the knowledge base.

As shown in Tab. 8, our method surpasses previous state-of-the-art methods, which demonstrates the effectiveness and generalization of our method. Tab. 9 shows the comparative results on the challenging A-OKVQA dataset. Our method achieved competitive results, which demonstrates the effectiveness of our method.