Soft Knowledge Prompt: Help External Knowledge Become a Better Teacher to Instruct LLM in Knowledge-based VQA

Qunbo Wang[1], **Ruyi Ji**[1], **Tianhao Peng**[2], **Wenjun Wu**[2], **Zechao Li**[3], **Jing Liu**[1]*
[1]Institute of Automation, Chinese Academy of Sciences, China
[2]Beihang University, China, [3]Nanjing University of Science and Techonolgy, China
{qunbo.wang,ruyi.ji}@ia.ac.cn, {pengtianhao,wwj09315}@buaa.edu.cn,
zechao.li@njust.edu.cn, jliu@nlpr.ia.ac.cn *

Abstract

LLM has achieved impressive performance on multi-modal tasks, which have received ever-increasing research attention. Recent research focuses on improving prediction performance and reliability (e.g., addressing the hallucination problem). They often prepend relevant external knowledge to the input text as an extra prompt. However, these methods would be affected by the noise in the knowledge and the context length limitation of LLM. In our work, we focus on making better use of external knowledge and propose a method to actively extract valuable information in the knowledge to produce the latent vector as a soft prompt, which is then fused with the image embedding to form a knowledge-enhanced context to instruct LLM. The experimental results on knowledge-based VQA benchmarks show that the proposed method enjoys better utilization of external knowledge and helps the model achieve better performance.

1 Introduction

Although LLM-based methods already perform well on many multi-modal tasks, knowledge-based VQA remains a challenging task, which requires outside knowledge beyond the image content to answer the question. In practice, LLM often generates factually incorrect responses to the given questions since the knowledge in the LLM model may be inaccurate, incomplete, and outdated. Intuitively, a straightforward solution is to incorporate relevant external knowledge to assist LLM in making a better prediction. In nature, such a paradigm is equivalent to using external knowledge as an additional prompt for LLM.

Related research efforts mainly focus on training a better retriever to obtain relevant external knowledge (Lin and Byrne, 2022), but the model passively receives the retrieved result. In fact, even the best retriever hardly guarantees that the retrieved results are without distracting information. Firstly, it is common that the retrieved knowledge might be irrelevant or part-irrelevant to answer the given question. Secondly, even if the retrieved knowledge is useful, LLMs often attend to the irrelevant parts or might completely ignore them, leading to generate the answer based on their incorrect knowledge. Therefore, it is non-trivial to study how to better utilize the retrieved knowledge.

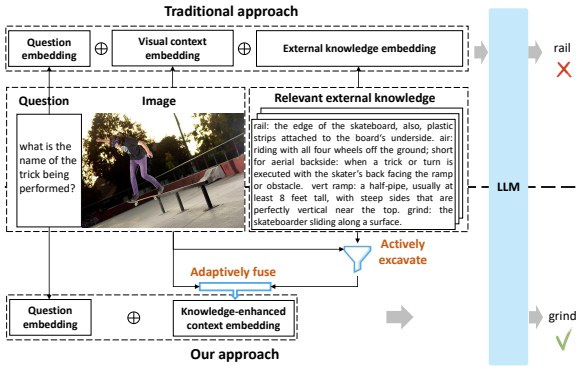

Figure 1: Comparison between the traditional approach and our approach to utilize relevant external knowledge. Giving the retrieved knowledge, the traditional approach directly prepends knowledge snippets to the input. In contrast, our approach dynamically selects and incorporates valuable information from the knowledge.

As shown in Figure 1, while containing supporting content, the external knowledge snippet also contains distracting content. The precisely supporting content renders a solid foundation for a model to generate the correct output. Various attempts have been made to filter retrieved snippets based on rules (Wang et al., 2023), which focus largely on the NLP tasks and neglect to excavate supporting information in a fine-grained manner. Currently, existing models fail to pay sufficient attention to the supporting content and are prone to be distracted by surrounding sentences that share similar topics (Shi et al., 2023).

*Jing Liu is the corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6132–6143

Multi-modal LLM (MLLM) has achieved tremendous success in multi-modal tasks. It leverages a projector to transform images into latent vectors that provide visual information context to the LLM. Given relevant external knowledge, prior approaches often directly prepend it with textual input to prompt LLM. In fact, the textual knowledge snippets can be analogously transformed into latent vectors to prompt LLM. Such behavior offers the following advantages: 1) realize the fine-grained extraction of pivotal information derived from external knowledge, facilitating the injection of knowledge into the prediction. 2) refine longer external knowledge texts into a small number of vectors, effectively avoiding adding numerous extra input tokens to LLM, which has a context length limitation.

Inspired by the empirical evidence mentioned above, we propose a novel method (coined SKP) to transform the external knowledge snippet into the latent vector as a **S**oft **K**nowledge **P**rompt to better instruct LLM for the knowledge-based VQA task. Specifically, we extract the pivotal information in the external knowledge according to the visual and question information and transform it into latent vectors, where the attention mechanism is introduced to select the informative tokens derived from the external knowledge. Then, knowledge latent vectors are fused with the visual embedding to form a knowledge-enhanced context for the question. Furthermore, we propose a tailored training scheme to expedite the process of selecting and fusing pivotal knowledge. For implementation details, please refer to the code at `https://github.com/BUAAw-ML/SKP.git`. The experimental results verify that the proposed method enjoys better utilization of external knowledge and helps the model achieve better performance. Our main contribution can be summarized into threefold:

- Different from other works, we dig into the solution for extracting and utilizing pivotal information from relevant knowledge to better instruct LLM, which can suppress distracting information and elicits knowledge into output.
- Differing from prepending external knowledge to the input, we embed the knowledge into a soft prompt and fuse it with the visual embedding, effectively avoiding adding extra input tokens to LLM.
- Experimental results on outside-knowledge based VQA datasets demonstrate that the proposed method achieves promising performance, validating the effectiveness of key design choices.

2 Related Work

2.1 Prompt LLM for Knowledge-based VQA

A huge amount of implicit knowledge is embedded in the parameters of pre-trained models, which are endowed with powerful knowledge capabilities when given proper prompts. Yang et al. (Yang et al., 2022) take image caption as a hard prompt to instruct GPT-3 for knowledge-based VQA. However, the generated captions fail to cover all the necessary information in an image. Some works utilize additional information to prompt models. For example, Rubin et al. (Rubin et al., 2022) retrieve related training examples as prompts for in-context learning of LLM. Xenos et al. (Xenos et al., 2023a) leverage question-informative captions and informative examples to prompt LLaMA. Shao et al. (Shao et al., 2023) further induce GPT-3 based on numerous candidate answers. Unlike hard prompts, Dai et al. (Dai et al.) project visual features into latent vectors to help small-scale LLMs perform better on knowledge-based VQA tasks, verifying the potential of soft prompts.

2.2 Retrieve Knowledge for Knowledge-based VQA

As the literature claims, external knowledge benefits answering questions when incorporated into visual-language models (Gardères et al., 2020; Zheng et al., 2021). Recently, retrieving textual knowledge snippets can help to achieve a better performance (Qu et al., 2021; Gao et al., 2022). Concretely, Luo et al. (Luo et al., 2021) retrieve knowledge in a cross-modal way. Gui et al. (Gui et al., 2022) construct a knowledge retriever on top of GPT-3. Lin and Byrne (Lin and Byrne, 2022) jointly optimizes T5 and retriever to excavate informative external knowledge. Lin et al. (Lin et al., 2023a) retrieve external knowledge by a fine-grained and multi-modal approach. Moreover, some research efforts construct knowledge embeddings for retrieval (Hu et al., 2023), which would incur and accumulate additional errors. Nevertheless, these methods vastly regard "Pseudo Relevance Labels" (Lin and Byrne, 2022) as training signals for retriever, which fail to retrieve sufficient pivotal

knowledge for a model to make correct prediction (Chen et al., 2022; Wang et al., 2024).

2.3 Improve External Knowledge Utilization

Recent research attempts, such as (Luo et al., 2023; Lin et al., 2023b), focus on better exploiting retrieved information based on instruction tuning LLM in the NLP domain. However, the requirements for LLM tuning hinder the wide application of such methods. Another research line focuses more on selecting useful external knowledge for LLM in the NLP domain. For instance, Ram et al. (Ram et al., 2023) train a reranker to select more informative external knowledge. Li et al. (Li et al., 2023) propose a verifiable generation method where LLM updates the retrieval result until meeting the supporting fact that the retrieved documents can benefit answer to the question. Baek et al. (Baek et al., 2023) fine-tune a small flan model to judge whether the retrieved knowledge is useful or not and recalibrate the knowledge engagement in the output. Some works select external knowledge to enhance the result generated by the model (Peng et al., 2023; Zhang et al., 2023). These works focus on selecting valuable knowledge from multiple external knowledge candidates. Differently, our method not only selects valuable knowledge but also makes more efforts to better extract the pivotal information from external knowledge.

Recently, Wang et al. (Wang et al., 2023) select valuable sentences from the knowledge fragment by matching or entropy mechanism without guaranteeing that the selected sentences contribute to making correct predictions. This work employs a method of filtering sentences that is coarse-grained, and it needs to train a filter, which does not guarantee ease of use and generalization. In addition, this work focuses on NLP tasks rather than multi-modal tasks. Existing methods struggle to extract pivotal knowledge in a fine-grained and accurate manner, especially in the multi-modal domain, rendering an open problem for the research community. Our attention-based method can help to mine the valuable information in a more fine-grained manner.

3 Method

In the MLLM framework, a projector module is often used to transform the image content into visual context vectors for LLM. We focus on extracting pivotal information from relevant external knowledge and fusing it with the visual context vectors as a knowledge-enhanced context to prompt LLM.

The framework of the proposed method is illustrated in Figure 2. Specifically, the input is well processed firstly (Sec. 3.1), and we propose a method to mine valuable information from relevant external knowledge (Sec. 3.2) and fuse it with the visual vectors to instruct LLM (Sec. 3.3). Further, we adopt an ensemble-based approach to determine the final output (Sec. 3.4). Lastly, we design a tailored optimization scheme for the trainable components involved in our framework (Sec. 3.5).

3.1 Input Representations

In our framework, there are three types of inputs: question, image, and relevant external knowledge.

Question embedding. For the input question text, we use the embedding function of LLM to obtain the question embedding vectors: $q \in \mathbb{R}^{n \times d}$, in which n is the total number of tokens, and d represents the embedded dimension.

Image embedding. For an image, we adopt ViT (Fang et al., 2023) to encode the image into visual embeddings. Further, following the common practice, the Q-Former module and the linear layer are used to transform visual embeddings into the latent space of LLM, in which t query tokens are implemented to extract visual information most relevant to the question. The output query representation serves as the visual representation, denoted as $V \in \mathbb{R}^{t \times d}$, which includes the visual information of interest to the question.

Knowledge embedding. In practice, external knowledge is often fed to the model as a text snippet. Therefore, we leverage the embedding function of LLM to obtain the knowledge snippet embedding $S \in \mathbb{R}^{l \times d}$, in which l is the token number of knowledge snippet text.

3.2 Knowledge Mining

This section proposes a method to extract pivotal knowledge information from relevant external knowledge. Firstly, our method selects top-k most valuable knowledge snippets from $k^{'}$ relevant external knowledge snippets ($k \leq k^{'}$). To obtain $k^{'}$ knowledge snippets per data sample, a public retriever *Vector Index Retriever* based on LlamaIndex (Liu, 2022) can be used. Then, regarding each selected knowledge, we extract the valuable information to produce the latent vector as a soft knowledge prompt.

Specifically, we use the visual representation V, which contains the visual information of interest

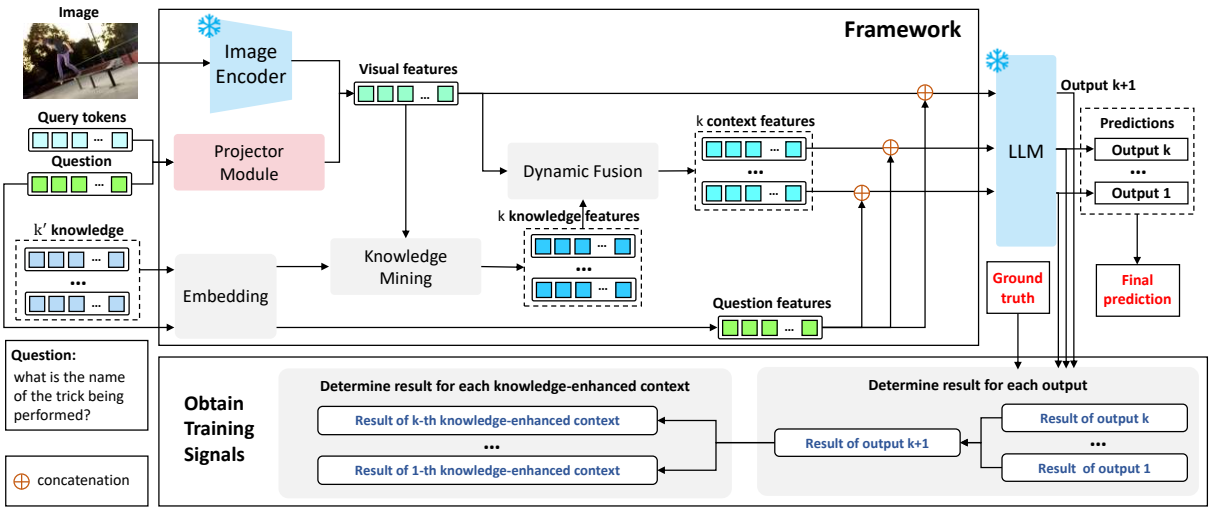

Figure 2: The framework of our method. Given relevant external knowledge, our method selects top-k valuable knowledge snippets and transforms them into vectors, which are fused with visual vectors to prompt LLM.

to the question, to help select valuable information in the external knowledge snippet. Firstly, we calculate the following value as a query for selecting valuable information:

$$v = \frac{1}{t}\sum_{i=1}^{t} V_i \tag{1}$$

where $v \in \mathbb{R}^d$. Because all the embeddings are in the latent space of LLM, we can directly measure the similarity between v and each token embedding $S_i (i \in [0, l])$ of the external knowledge snippet. To capture more detailed relations, we calculate the multi-head similarity between v and S_i. Specifically, v and S_i are firstly multiplied with the parameter matrices W_1 and W_2 respectively:

$$\begin{aligned} v^{'} &= W_1 v \\ S_i^{'} &= W_2 S_i \end{aligned} \tag{2}$$

where $W_1 \in \mathbb{R}^{d\times d}$ and $W_2 \in \mathbb{R}^{d\times d}$. Then, for the two projected vectors $v^{'}$ and $S_i^{'}$, the last dimension of the vectors is split from $(..., d)$ to $(..., h, d/h)$. Thus, we can transform each vector into h vectors:

$$\begin{aligned} v^{'} &\rightarrow \{v_1^{'}, \ldots, v_h^{'}\} \\ S_i^{'} &\rightarrow \{S_{i1}^{'}, \ldots, S_{ih}^{'}\} \end{aligned} \tag{3}$$

where h is the number of heads. Note that we set its value equal to the number of query tokens t ($h = t$). Lastly, for each pair (e.g. $v_1^{'}$ and $S_{i1}^{'}$), we calculate the dot product similarity ϕ^{dot} between them, and we can obtain h similarity values:

$$\{\phi^{dot}(v_1^{'}, S_{i1}^{'}), ..., \phi^{dot}(v_h^{'}, S_{ih}^{'})\} \tag{4}$$

For l tokens in an external knowledge snippet, we calculate the multi-head similarity of each token according to the above process individually. Therefore, the dimension of the similarity value of an external knowledge is $l \times h$. And we denote this similarity value as $R \in \mathbb{R}^{l\times h}$. Further, to obtain a score representing the external knowledge similarity, we take the mean of the similarity values of the l tokens as the final similarity score of an external knowledge text, which can be denoted as the following equation:

$$r = \frac{1}{l}\sum_{i}^{l}[\frac{1}{h}\sum_{j}^{h}(R_{ij})] \tag{5}$$

Based on the similarity score r, we select top-k knowledge snippets from $k^{'}$ knowledge snippets.

After acquiring the selected knowledge snippet, we propose an attention-based method to extract pivotal information and form a knowledge representation. Specifically, based on the relevance matrix $R \in \mathbb{R}^{l\times h}$, we calculate the relevance score of each token for h heads respectively, where the SoftMax function is introduced to obtain the normalized relevance score for each head $R_j^{'} \in \mathbb{R}^l (j \in [1, h])$. Then, according to the attention-based mechanism, we use the normalized relevance score $R_j^{'}$ as the weight vector and compute the weighted sum of the knowledge tokens through the following equation:

$$P_j = \frac{1}{l}\sum_{i=1}^{l} R'_{ji} S_i \tag{6}$$

where $P_j \in \mathbb{R}^d (j \in [1, h])$.

Last, the knowledge representations of all heads are concatenated together as a soft knowledge prompt:

$$P = [P_1; ...; P_h] \tag{7}$$

where $[.;.]$ denotes the concatenation operation, and the soft knowledge prompt is denoted as $P \in \mathbb{R}^{h \times d}$.

3.3 Dynamic Fusion

The assumption hardly holds that the retriever always acquires useful knowledge. Besides, the importance of images and external knowledge for answering the question stochastically fluctuates for various samples. Therefore, after obtaining the soft knowledge prompt P, we introduce an adaptive fusion mechanism to dynamically fuse P with the visual representation V to obtain a knowledge-enhanced context to better instruct LLM.

To balance the proportion of external knowledge information and visual information, we concatenate V and P to predict scalars for fusion. Specifically, we use a gate-like structure with a Softmax function to dynamically calculate the weight value:

$$[\sigma_1, \sigma_2] = \mathrm{Softmax}([\mathrm{FFN}_1(\mathrm{V}); \mathrm{FFN}_2(\mathrm{P})]) \tag{8}$$

where $\sigma_1 \in \mathbb{R}^{t \times 1}$ and $\sigma_2 \in \mathbb{R}^{h \times 1}$. Note that $t = h$. Then, based on the fusion weights, we employ the weighted sum of V and P to generate the knowledge-enhanced context $C \in \mathbb{R}^{t \times d}$:

$$C = \sigma_1 V + \sigma_2 P \tag{9}$$

3.4 Answer Generation

Conditioned on k relevant external knowledge snippets, we can obtain k knowledge-enhanced contexts. Then, we concatenate each knowledge-enhanced context with the input question text and feed it into LLM to produce an answer output o. Then, majority voting is performed over the k output answers to determine the answer res:

$$res = \arg\max\left(\mathrm{count}(o_1, \ldots, o_k)\right) \tag{10}$$

where the function $\mathrm{count}()$ records occurrences of each answer in the k outputs.

3.5 Optimization

In our design, the visual encoder and LLM parameters are frozen, and other parameters are trained to help better excavate pivotal information from external knowledge. Hence, the optimization scheme should encourage the occurrence of knowledge-enhanced contexts that are truly useful for prediction, and avoid knowledge-enhanced contexts without any contribution to the model's prediction.

Firstly, we need to identify which knowledge-enhanced contexts are truly useful. Intelligibly, the positive knowledge-enhanced context should help the model make correct predictions when the model predicts incorrectly without using the knowledge. To achieve this goal, the input using only the image and question is fed into LLM and the output is denoted as o_{k+1}. Then, a knowledge-enhanced context can be judged as positive when the LLM output using this knowledge-enhanced context is true and the output o_{k+1} is wrong. In summary, we identify three subsets of the knowledge-enhanced context based on whether they can correct the answer:

$$\begin{aligned} R^P &= \{i \in [1,k] : o_i = o^* \wedge o_{k+1} \neq o^*\} \\ R^N &= \{i \in [1,k] : o_i \neq o^* \wedge o_{k+1} = o^*\} \\ R^I &= \{i \in [1,k] : i \notin R^P \wedge i \notin R^N\} \end{aligned} \tag{11}$$

where o^* denotes the ground truth answer, R^P denotes the set including the positive knowledge-enhanced context, and R^N denotes the set covering the negative knowledge-enhanced context that confuses answer generation. Furthermore, we denote R^I as the set that belongs to neither R^P nor R^N, which is absent from calculating loss during training. Here our goal is to decrease the similarity scores of snippets in R^N, and increase the similarity scores of snippets in R^P. Mathematically, the model optimization is constricted by the following loss:

$$-\sum_{(x,o*)\in TrainSet}[\sum_{i} log\mathrm{p}(o^*|x,C_i) + \sum_{i\in R^P} log\mathrm{p}(C_i|x) - \sum_{i\in R^N} log\mathrm{p}(C_i|x)] \quad (12)$$

where x denotes the input sample. The first term improves the answer generation performance. The second and last terms affect the external knowledge utilizing.

4 Experiments

4.1 Datasets

OK-VQA. The OK-VQA dataset (Marino et al., 2019) is a widely-used dataset for the outside-knowledge based VQA task, evaluating related knowledge-enhanced models. The questions in this dataset are crowd-sourced from Amazon Mechanical Turkers (AMT) [1]. The dataset consists of 14,055 questions, of which $9,009$ questions are held out for training and $5,046$ questions are held out for testing. In addition, GS knowledge corpus (Luo et al., 2021) is used by the retriever to obtain the relevant knowledge in our experiments, which is a benchmark knowledge source for OK-VQA.

A-OKVQA. The A-OKVQA dataset (Schwenk et al., 2022) is a crowd-sourced dataset, which requires a wide range of world and commonsense knowledge to answer the questions. This dataset contains about 25K questions, in which 17K questions are used for training, and each sample on the training and validation sets is supplied with three relevant external knowledge snippets.

4.2 Experimental Setup

To thoroughly compare our results against existing methods, various evaluation metrics are adopted to measure the model performance. Specifically, VQA Score is the most used metric in VQA task, which is calculated with pre-processed human annotations A:

$$VQAScore(y,A) = \min(\frac{\#A(y)}{3}, 1) \quad (13)$$

in which $\#A(y)$ denotes the number of annotators who answered y.

Exact Match (EM) is also adopted as the evaluation metric in our work, which treats annotated answers equally:

$$EM(y,A) = \min(\#A(y), 1) \quad (14)$$

[1]https://www.mturk.com

4.3 Implementation Details

We optimize the proposed model in the Pytorch framework with AdamW optimizer, a batch size of 1, gradient accumulation step of 8, and we use NVIDIA A6000 GPUs. The initial learning rate is set to 1e-5, and the weight decay is set to 0.05; the warm-up learning rate and steps are set to 1e-8 and 1000, respectively. In our method, t is set to 32, and h is also set to 32. In addition, $k^{'}$ is set to 10 in experiments of OK-VQA and set to 3 in experiments of A-OKVQA. The experiment of each method is repeated 5 times, and the final result is calculated via the average over the 5 runs.

4.4 Comparison Methods

On the OK-VQA dataset, we compare the proposed method with a wide range of public approaches. Wherein, ConceptBERT (Gardères et al., 2020), KRISP (Marino et al., 2021), UnifER (Guo et al., 2022), VRR (Luo et al., 2021), KAT-T5 (Gui et al., 2022), TRiG-Ensemble (Gao et al., 2022), RA-VQA (Lin and Byrne, 2022), MAVEx (Wu et al., 2022), RA-VQA-v2 (Lin et al., 2023a), VLC-BERT (Ravi et al., 2023) and TRiG (Ensemble) (Gao et al., 2022) are relatively small in model size (<10B), MM-Reasoner (Khademi et al., 2023), PICa (Yang et al., 2022), KAT (Gui et al., 2022), Prophet (Shao et al., 2023), PromptCap (Hu et al., 2022), REVIVE (Lin et al., 2022), Flamingo (Alayrac et al., 2022), and TwO (Si et al., 2023) use very large models such as GPT-3 (175B). Note that (Sun et al., 2023) and (Xenos et al., 2023b) do not offer the name of the proposed method.

On the A-OKVQA dataset, we compare with various methods, including ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), ClipCap (Mokady et al., 2021), KRISP, GPV-2 (Kamath et al., 2022), Prophet (Shao et al., 2023).

4.5 Main Results

Table 1 reports the comparison results on the OK-VQA dataset. One can figure out the following observations. Firstly, compared with the methods (e.g., KAT-T5, RA-VQA) retrieving textual knowledge snippets, knowledge graph-based methods (e.g., ConceptBERT, KRISP) perform less satisfactorily. The reason may be that knowledge graph-based methods fail to gain enough knowledge for the outside-knowledge question-answering task.

Method	Base Models	k	EM	VQAScore
ConceptBERT (Gardères et al., 2020)				33.7
KRISP (Marino et al., 2021)				38.4
Visual Retriever-Reader (Luo et al., 2021)		100		39.2
MAVEx (Wu et al., 2022)				39.4
UnifER (Guo et al., 2022)				42.1
VLC-BERT (Ravi et al., 2023)				43.1
KAT-T5 (Gui et al., 2022)	T5-large(770M)	40		44.3
(Sun et al., 2023)				49.4
TRiG (Gao et al., 2022)	T5-large(770M)	100	54.7	50.5
RA-VQA (Lin and Byrne, 2022)	T5-large(770M)	50	59.4	54.5
RA-VQA-v2 (Lin et al., 2023a)	T5-XL(3B)	5	62.0	62.1
Method based on large-scale model (>10B parameters)				
PICa (Yang et al., 2022)	GPT-3 (175B)			48.0
KAT (Single) (Gui et al., 2022)	T5-large (770M)+ GPT-3 (175B)			53.1
KAT (Ensemble) (Gui et al., 2022)	T5-large (770M)+ GPT-3 (175B)	40		54.4
REVIVE (Single) (Lin et al., 2022)	GPT-3 (175B)			56.6
Flamingo (Alayrac et al., 2022)	Flamingo(80B)			57.8
REVIVE (Ensemble) (Lin et al., 2022)	GPT-3 (175B)			58.0
TwO (Si et al., 2023)	GPT-3 (175B)			58.7
PromptCap (Hu et al., 2022)	GPT-3 (175B)			60.4
MM-Reasoner (Khademi et al., 2023)	GPT-4 + i-Code v2			60.4
Prophet (Shao et al., 2023)	GPT-3 (175B) +MCAN			61.1
(Xenos et al., 2023b)	LLaMA 2 (13B) +MCAN			61.2
Our proposed method and ablation versions				
SKP (FlanT5)	FlanT5-XXL	3	61.1	56.4
w/ traditional knowledge incorporation approach	FlanT5-XXL	3	59.7	54.9
SKP (vicunna)	vicunna-7b	3	**68.9**	**63.3**
w/ traditional knowledge incorporation approach	vicunna-7b	3	67.0	61.2
w/o dynamic fusion	vicunna-7b	3	67.6	62.4
w/o knowledge supervision signal	vicunna-7b	3	68.3	62.9
w/o ensemble-based prediction	vicunna-7b		68.7	63.0

Table 1: Performance on OK-VQA dataset. k denotes the number of knowledge snippets used for prediction. The best performance is bolded. Given the same model and knowledge, our method performs better than the traditional knowledge incorporation approach, indicating that it can better mine the information in external knowledge.

Secondly, the methods based on the large-scale LLM (e.g., GPT-3) can achieve good performance. And we can find the methods using better prompts (e.g., Prophet) that can induce the same LLM to achieve better performance. Lastly, our method, transforming the external knowledge into embeddings as a soft prompt, achieves decent performance even with a small-scale LLM.

We further conduct ablation studies on OK-VQA. As listed in Table 1, with regard to the traditional knowledge incorporation approach, we observe a significant drop in performance, which indicates that transforming the external knowledge into a soft prompt improves the instruction of LLM. As for the variant without dynamic fusion, denoted as SKP (w/o dynamic fusion), we see that the performance degrades from 63.3 to 62.4. We conjecture that dynamic fusion adaptively attends to external knowledge, mitigating the influence of cases where all knowledge snippets are irrelevant. Moreover, we perform an experiment without the knowledge supervision signal, denoted as SKP (w/o knowledge supervision signal), and a performance decline is observed, which manifests the effectiveness of the proposed training method in terms of knowledge utilization. And this supervision signal can play a more important role as more knowledge snippets are selected for answer generation because it inevitably introduces more noise and increases the need for extracting pivotal information. In addition, when the variant work without an ensemble-based prediction mechanism, denoted as SKP (w/o ensemble-based prediction), we notice a drop in performance, which validates the necessity of multiple external knowledge. We also verified the significance of design choices with

Model	VQAScore
ViLBERT (Lu et al., 2019)	30.6
LXMERT (Tan and Bansal, 2019)	30.7
ClipCap (Mokady et al., 2021)	30.9
KRISP (Marino et al., 2021)	33.7
GPV-2 (Kamath et al., 2022)	48.6
Prophet (Shao et al., 2023)	58.2
SKP (vicunna-7b)	65.3
w/ traditional knowledge incorporation approach	63.8

Table 2: Experimental results on A-OKVQA.

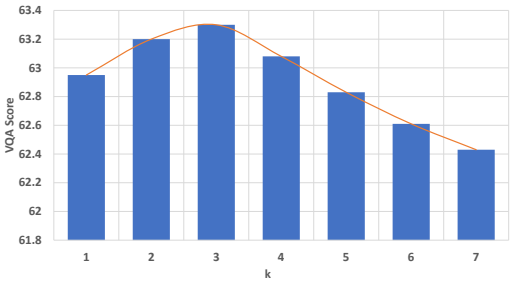

Figure 3: Analysis of the parameter k.

$\text{scipy.stats.ttest\_ind}(\text{p} < 0.05)$.

On the other hand, we conduct experiments on the A-OKVQA dataset, and k is set to 3. The results are obtained based on the direct answer evaluation setting, which is more challenging and realistic. In addition, because our work aims to better utilize relevant knowledge and A-OKVQA provides the relevant knowledge on the training and validation sets, we provide the comparison results on the validation set. As summarized in Table 2, it can be seen that our method performs better than the others on the A-OKVQA dataset. Notably, the performance drops dramatically when engaged with the traditional knowledge incorporation approach. We attribute this to the fact that our design choice endows our model with the robust capability to handle external knowledge.

4.6 Effects of k

This section investigates how our method behaves using a different number of knowledge snippets k on the OK-VQA dataset. Firstly, we set $k = \{1, 2, 3, 4, 5, 6, 7\}$ to evaluate model performance respectively. As Figure 3 shows, the prediction performance fluctuates as k varies and saturates with $k = 3$. We conjecture that the prediction under the consideration of multiple external knowledge has a positive effect, but when too much relevant knowledge is involved (k gradually approaches $k^{'}$), the knowledge choice space for the proposed method greatly shrinks, and more distracting information messes pivotal knowledge up, leading to the limited performance. Therefore, we empirically set $k = 3$ and $k^{'} = 10$ in our experiments.

4.7 Case Studies

Figure 4 exhibits two samples of prediction. Specifically, in the first question, "what type of apples is the woman using in this recipe?", it is obvious that the question needs the help of external knowledge to make an accurate answer (instead of a superficial answer), in which using only the implicit knowledge of LLM does not perform well. Besides, conditioned on the traditional knowledge incorporation method, it remains challenging to generate a correct answer. We conjecture that the traditional approach fails to capture the information derived from external knowledge accurately. In the second question, "which type of metal is used for making this toilet? ", the answer obtained conditioned on LLM only deviates from the essence of the actual situation. That is, the toilet itself cannot be made of aluminum. So, the wrong answer is caused by the incomplete knowledge contained in LLM, or the accurate knowledge of LLM is not induced. Further, the answer obtained by the traditional knowledge-enhanced method is still not accurate enough, indicating that this paradigm lacks the ability to handle external knowledge accurately.

4.8 Discussions

The prompt method is crucial for eliciting the knowledge of large language models in the output. In our experiments, the proposed prompt method outperforms the 'SKP(vicunna) w/ traditional knowledge incorporation approach', which concatenates external knowledge text and image embedding based on projector. This demonstrates that our prompt method can better stimulate the performance of LLM on downstream tasks. Additionally, we conducted an experiment on OK-VQA using a typical prompt method that converts the image into textual descriptions instead of using Q-former to transform them into embeddings. The results were very poor (below 50) because this prompt method cannot ensure that all image information is captured and may miss visual details relevant to the question.

Compared to the outside-knowledge based datasets OK-VQA and A-OKVQA, FVQA requires basic factual knowledge to answer questions. We

Question: what type of apples is the woman using in this recipe? (ground-truth answer: granny smith)

Answer without using external knowledge: **green**
Answer using external knowledge based on the traditional method: **green**
Answer using external knowledge based on the proposed method: **granny smith**

Demonstration of words with the attention weights in one external knowledge using our method:

the	granny	smith	apple	variety	is	unique	in	the	apple
world	with	its	green	colour	and	very	sour	taste	,
making	it	the	exact	counter	part	to	the	golden	delicious

Question: which type of metal is used for making this toilet? (ground-truth answer: stainless steel)

Answer without using external knowledge: **aluminum**
Answer using external knowledge based on the traditional method: **stainless**
Answer using external knowledge based on our method: **stainless steel**

Demonstration of words with the attention weights in one external knowledge using our method:

some	kind	of	lavatory	flushed	with	water	is	believed	to
have	been	used	by	residents	of	the	indus	valley	by
around	2000	b.c	.	toilet	seats	are	generally	made	from
one	of	two	materials	.	the	metal	tank	fixtures	are
made	of	stainless	steel	or	copper	,	and	the	joints
that	hold	the	seat	to	the	bowl	are	usually	a
rubber	-	like	plastic	.	or	used	it	as	a
urinal	and	then	the	""	slop	jar	""	was	emptied
into	the	outhouse							

Figure 4: Two samples on OK-VQA. Based on the same model, the results without using external knowledge, using external knowledge based on the traditional method, and using external knowledge based on our method are given.

conducted an experiment on this dataset and found that our method, given the same external knowledge, enables the model to achieve higher performance than traditional methods. However, because the questions in FVQA often require relatively simple external knowledge, the implicit knowledge in LLMs is already sufficient to handle these questions. As a result, utilizing external knowledge does not significantly improve performance.

5 Conclusion

This work proposes a method to extract the pivotal information from the relevant external knowledge, which is then fused with the visual information to form a knowledge-enhanced context. This approach protects LLM from being misled by distracting information, and avoids adding extra input tokens to LLM. The experimental results on knowledge-based VQA benchmarks demonstrate that the proposed method enjoys better utilization of external knowledge and is conducive to more accurate answers. In addition, compared with constructing a large and complex knowledge mining model, which requires lots of labeled training data and faces the problem of generalization, our method enjoys the characteristics of good interpretability, and simplicity of use.

Different from most recent works, we do not focus on constructing a better retriever to obtain more beneficial knowledge. Given a relevant external knowledge, our work focuses on how to better utilize it to improve the prediction performance. We try our best to excavate its value, but the beneficial information within an external knowledge is limited. In future work, we will study how to combine our method and the retriever construction method to obtain a better performance improvement. In addition, we will also investigate the potential to improve the complementary between external knowledge and model implicit knowledge.

Limitations

Our method focuses on improving the prediction performance of LLM by mining external knowledge, but there is vast implicit knowledge embedded in LLM. Therefore, the problem of how external knowledge and implicit knowledge supplement each other mutually remains tricky. In the case where external knowledge and implicit knowledge conflict, it is necessary for a model to infer the quality of knowledge. The above problems are beyond the scope of this paper and left for research efforts in the future.

Acknowledgements

This paper is supported by National Key R&D Program of China (No. 2022ZD0118802), National Natural Science Foundation of China (No. 62206279, 62306314), Jiangsu Key R&D Program for Industry Prospect and Core Technological Innovations Project (No. BE2023016).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C Park, and Sung Ju Hwang. 2023. Knowledge-augmented language model verification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Chongqing Chen, Dezhi Han, and Chin-Chen Chang. 2022. Caan: Context-aware attention network for visual question answering. *Pattern Recognition*, 132:108980.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

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.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968.

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.

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

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379.

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.

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.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023. Llatrieval: Llm-verified retrieval for verifiable generation. *arXiv preprint arXiv:2311.07838*.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023a. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023b. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

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.

Jerry Liu. 2022. LlamaIndex.

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.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431.

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.

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

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1753–1757.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. Vlc-bert: visual question answering with contextualized commonsense knowledge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1155–1165.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

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*, pages 146–162. Springer.

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.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of thinking and observing for outside-knowledge VQA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10959–10975. Association for Computational Linguistics.

Zhongfan Sun, Yongli Hu, Qingqing Gao, Huajie Jiang, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2023. Breaking the barrier between pre-training and fine-tuning: A hybrid prompting model for knowledge-based vqa. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4065–4073.

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*, pages 5100–5111.

Qunbo Wang, Jing Liu, and Wenjun Wu. 2024. Coordinating explicit and implicit knowledge for knowledge-based vqa. *Pattern Recognition*, page 110368.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

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. 2023a. A simple baseline for knowledge-based visual question answering. *arXiv preprint arXiv:2310.13570*.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023b. 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, Singapore. Association for Computational Linguistics.

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.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain qa. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728.

Wenfeng Zheng, Lirong Yin, Xiaobing Chen, Zhiyang Ma, Shan Liu, and Bo Yang. 2021. Knowledge base graph embedding module design for visual question answering model. *Pattern recognition*, 120:108153.