Knowledge Generation for Zero-shot Knowledge-based VQA

Rui Cao and **Jing Jiang**
School of Computing and Information Systems
Singapore Management University
ruicao.2020@phdcs.smu.edu.sg, jingjiang@smu.edu.sg

Abstract

Previous solutions to knowledge-based visual question answering (K-VQA) retrieve knowledge from external knowledge bases and use supervised learning to train the K-VQA model. Recently pre-trained LLMs have been used as both a knowledge source and a zero-shot QA model for K-VQA and demonstrated promising results. However, these recent methods do not explicitly show the knowledge needed to answer the questions and thus lack interpretability. Inspired by recent work on knowledge generation from LLMs for text-based QA, in this work we propose and test a similar knowledge-generation-based K-VQA method, which first generates knowledge from an LLM and then incorporates the generated knowledge for K-VQA in a zero-shot manner. We evaluate our method on two K-VQA benchmarks and found that our method performs better than previous zero-shot K-VQA methods and our generated knowledge is generally relevant and helpful. [1]

1 Introduction

Knowledge-based VQA (which we refer to as K-VQA in this paper) is a special visual question answering (VQA) task where, in addition to an image, external knowledge is needed to answer the given question. For instance, to answer the question in Figure 1, background knowledge about national parks in California is needed.

Early methods for K-VQA follow a *retrieve and answer* paradigm (Figure 1(a)), which first retrieves knowledge from external knowledge sources as additional input and then trains a VQA model through supervised learning (Wang et al., 2018; Narasimhan and Schwing, 2018; Narasimhan et al., 2018; Li et al., 2020). This paradigm requires both a suitable external knowledge base and a large amount of K-VQA training data, which may not be practical for real applications when either of these resources is not available. Recently, with the fast advances of LLMs that have demonstrated remarkable zero-shot transfer capabilities, several studies applied LLMs for K-VQA under zero-shot or few-shot settings, leveraging both the extensive knowledge implicitly contained in LLMs and their built-in question answering capability (Yang et al., 2022; Hu et al., 2022; Guo et al., 2022; Li et al., 2023a; Alayrac et al., 2022). Typically, these methods first convert an image to text descriptions (i.e., captions) and then feed the captions and the question into an LLM to directly obtain the answer, as illustrated as the *directly answer* paradigm in Figure 1(b).

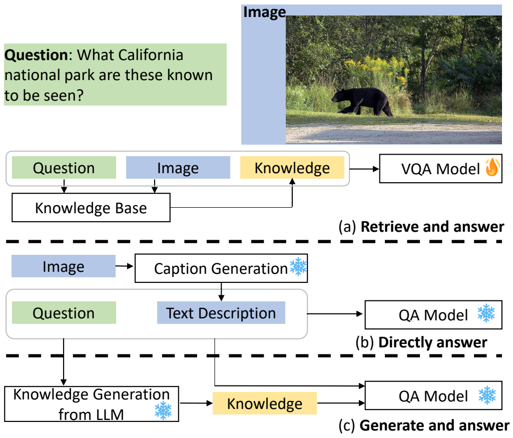


Figure 1: Three approaches to K-VQA: retrieve and answer, directly answer, and generate and answer.

However, none of these zero-shot or few-shot methods *explicitly* states the knowledge needed to answer a question. As we know, answering K-VQA questions usually requires external knowledge not seen in the image. Even if the external knowledge is implicitly contained in the LLM used for QA, it is not immediately clear whether and how the LLM can use the relevant knowledge to answer a K-VQA question through the *directly answer* paradigm. On the other hand, recent work has shown that for text-based QA that requires multi-step reasoning, explicitly generating relevant knowledge and including it as additional input improves QA performance (Liu

[1] Code available: https://github.com/abril4416/KGen_VQA

Findings of the Association for Computational Linguistics: EACL 2024, pages 533–549
March 17-22, 2024 ©2024 Association for Computational Linguistics

et al., 2022; Yu et al., 2023). We suspect that this is also the case for K-VQA. Furthermore, explicitly generated knowledge improves the explainability of the system. Another limitation of previous zero-shot and few-shot K-VQA methods is that some of them rely on task-specific training such as the training of a question-specific caption generation model in PromptCap (Hu et al., 2022), which still requires significant amount of training data.

In this paper, we attempt to address these limitations of previous work. Inspired by Liu et al. (2022), which uses an LLM to generate explicit knowledge statements to facilitate text-based commonsense QA, we propose a similar zero-shot K-VQA method that uses an LLM (specifically GPT-3) to *explicitly* generate potentially useful knowledge statements to facilitate K-VQA, as illustrated in Figure 1(c). In addition to having explicit knowledge statements, our method is also free from any additional training. To improve the diversity and coverage of the generated knowledge, we further borrow the self-supervised knowledge diversification strategy from (Yu et al., 2023). We call our method KGenVQA. To the best of our knowledge, we are the first to test the *generate and answer* approach on K-VQA.

We evaluate KGenVQA on both OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), two benchmark datasets commonly used for K-VQA. The experiments demonstrate that our generated knowledge statements are effective in improving the K-VQA performance in terms of answer accuracy, when everything else being equal, and our method can outperform SOTA zero-shot K-VQA methods that do not use extra training. We also measure the usefulness of our generated knowledge and find that the generated knowledge statements have high quality in terms of grammaticality, relevance, factuality, helpfulness, and diversity, based on manual judgement. Our findings demonstrate that *generate and answer* is a feasible zero-shot approach to K-VQA with the additional benefit of providing explanations through the explicitly generated knowledge statements.

2 Related Work

K-VQA. Early K-VQA models were built through standard supervised training, with a large amount of (Image, Question, Answer) triplets as training data (Wang et al., 2018; Narasimhan and Schwing, 2018; Narasimhan et al., 2018; Li et al., 2020). Typically, these models retrieve knowledge from an external knowledge source such as ConceptNet or Wikipedia and use the retrieved knowledge to facilitate QA. In our work, we also use explicit knowledge to facilitate QA, but the knowledge is generated from an LLM instead.

Zero-shot K-VQA. Several recent studies utilized LLMs for zero-shot K-VQA (Yang et al., 2022; Hu et al., 2022; Guo et al., 2022; Li et al., 2023a; Alayrac et al., 2022). Generally, these methods first convert the given image into captions or embeddings compatible with a pre-trained language model. Then the captions or embeddings are combined with the question as input to the language model for zero-shot QA. We can categorize these methods into two types: those that need extra training using labeled data other than K-VQA data, and those that directly leverage existing pre-trained models without any further training or fine-tuning. Examples of the former category include Frozen (Tsimpoukelli et al., 2021) (which uses image-text pairs to train a projection module) and BLIP-2 (Li et al., 2023a) (which learns a Q-transformer module to model multimodal interactions). Examples of the latter category include PICa (Yang et al., 2022) and PNP-VQA (Tiong et al., 2022), which convert the images into captions with an off-the-shelf caption generator. However, to the best of our knowledge, none of the existing zero-shot K-VQA methods explicitly state the external knowledge used to answer the questions.

Knowledge generation for QA. A few recent studies on text-based QA tested the idea of using LLMs to generate either short knowledge statements or long documents before combining them with the questions for zero-shot commonsense QA or open-domain QA (Liu et al., 2022; Sun et al., 2022; Yu et al., 2023). They found that by incorporating the generated knowledge in QA, performance can be significantly improved. Our work is inspired by these recent studies but we apply the idea to visual QA.

3 Method

The high-level idea of our KGenVQA method is to leverage an LLM to generate explicit knowledge statements given an image and a question. These knowledge statements can then be combined with the image captions and the question to be passed to the same or a different LLM for zero-shot text-based QA. In this section, we first elaborate how we generate knowledge statements from an LLM using few-shot in-context learning. We then present how the generated knowledge is integrated into the question answering process.

3.1 Knowledge Generation

Our knowledge generation process consists of two steps: An *initial knowledge generation* step, in which we generate a single knowledge statement for each (image, question) pair in the K-VQA test dataset, and a subsequent *self-supervised knowledge diversification* step, in which we sample a diverse set of knowledge statements generated during the first step as in-context demonstrations to perform a second round of knowledge generation, in which we generate multiple knowledge statements per (image, question) pair. The motivation is that with a diverse set of in-context demonstrations, we expect the LLM to also generate knowledge statements covering different aspects of the same (image, question) pair, which may increase the chance of getting the correct answer.

Caption generation. In both knowledge generation steps, we regard an LLM (GPT-3 in our experiments) as a knowledge base because the LLM has been trained on a large amount of text covering a wide range of topics. Previous work has shown that relevant knowledge statements can be generated from an LLM if appropriate text prompts including both the contexts and some demonstrations are used (Liu et al., 2022). However, different from text-based QA, for K-VQA, the context is an image, which cannot be directly used as input to an LLM. To address this issue, we adopt a simple solution that converts the image into one or more captions, using an off-the-shelf image captioning model. However, instead of using a general-purpose captioning model, we believe that *question-aware* captions, which focus on describing the parts of the image that are more relevant to the question, can provide better contexts for knowledge generation. Therefore, we adopt the question-aware caption generation mechanism by Tiong et al. (2022), which first highlights image regions that are more relevant to the question and then generates question-aware captions with the attention-weighted image. Following the practice of Tiong et al. (2022), we use multiple captions because this practice has been shown to be useful for subsequent question answering. We concatenate the multiple captions into a single sequence of tokens, which we denote as C.

Prompt template for knowledge generation. In both the initial knowledge generation step and the knowledge diversification step, to generate a single piece of knowledge, we use the following prompt template: *Please generate related background knowledge to the question*; *Context:* $[C]$; *Question:* $[Q]$; *Knowledge:*. The LLM will complete the prompt above by generating a sentence, which we treat as a knowledge statement. In order to better generate the relevant knowledge, we leverage in-context learning by including a few demonstrations, i.e., a few examples each containing a context (which are also image captions), a question, and the expected knowledge statement to be generated. During the initial knowledge generation step and the knowledge diversification step, we use different kinds of demonstrations.

Initial knowledge generation. During the initial knowledge generation step, we use six manually crafted in-context demonstrations for knowledge generation. They can be found in Appendix H. During this step, we generate a single knowledge statement for each (image, question) pair in a K-VQA test dataset.

Self-supervised knowledge diversification. Previous work showed that proper selection of demonstrations is of vital importance when prompting LLMs (Yang et al., 2022; Gonen et al., 2022). We suspect that the manually crafted demonstrations may not always be proper examples for all test instances. Besides, when answering knowledge-intensive questions, oftentimes more than one piece of knowledge may be needed. For instance, to answer the question in Figure 2, the knowledge 1) what national parks are in California; 2) among national parks in California, which is famous for black bears. To generate multiple knowledge statements per question, a straightforward solution is to ask the LLM to return multiple pieces of knowledge. However, beam search sampling, as mentioned in (Holtzman et al., 2020; Vijayakumar et al., 2018), tends to generate dull and repetitive outputs, and the improved top-k sampling (Fan et al., 2018) can only solve the issue to some extent. On the other hand, with different prompts, an LLM may generate diverse outputs (Li et al., 2023b).

Therefore, we adopt a self-supervised knowl-

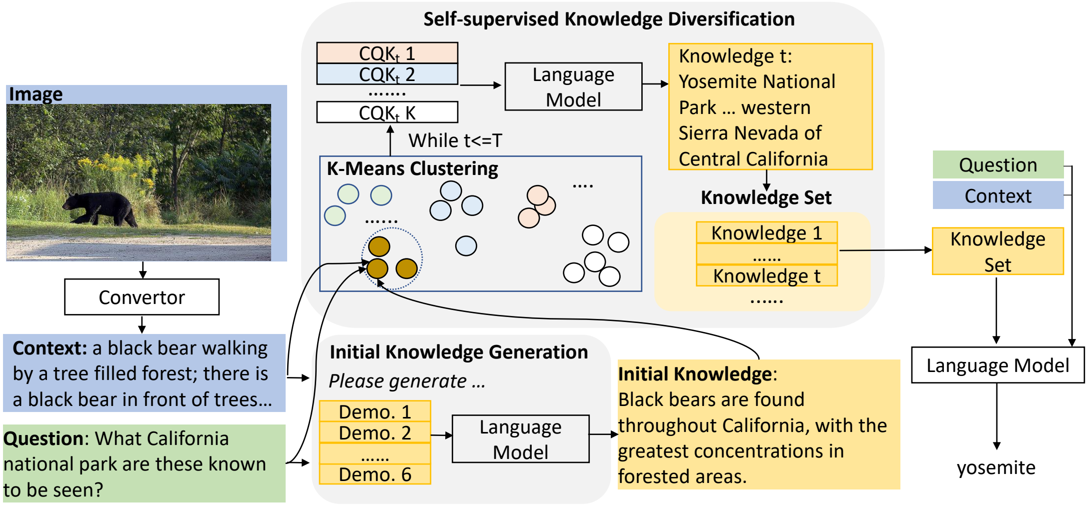

Self-supervised Knowledge Diversification
CQK_t 1
CQK_t 2
....
CQK_t K
While t<=T
Language Model
Knowledge t: Yosemite National Park ... western Sierra Nevada of Central California
Knowledge Set
K-Means Clustering
......
....
Knowledge 1
......
Knowledge t
......
Question
Context
Knowledge Set
Language Model
yosemite
Image
Convertor
Context: a black bear walking by a tree filled forest; there is a black bear in front of trees...
Question: What California national park are these known to be seen?
Initial Knowledge Generation
Please generate ...
Demo. 1
Demo. 2
......
Demo. 6
Language Model
Initial Knowledge: Black bears are found throughout California, with the greatest concentrations in forested areas.

Figure 2: An overview of the proposed method. We first convert the image into textual descriptions and prompt LLMs with the question and manual demonstrations to obtain the initial knowledge pieces. In the second stage, we diversify knowledge by selecting a diverse set of knowledge statements in the first step as demonstrations. Lastly, we incoporate the generated knowledge for QA with a language model.

edge diversification strategy by (Yu et al., 2023) as follows. Let $\mathcal{K}_{\text{init}} = \{(C_i, Q_i, K_i)\}_{i=1}^{N}$ denote the set of (captions, question, knowledge statement) triplets obtained during the initial knowledge generation step, where K_i is the knowledge statement generated for (C_i, Q_i). We treat each triplet (C_i, Q_i, K_i) as a "silver"-labeled demonstrating example. Slightly different from (Yu et al., 2023), we hypothesize that if each time we sample a different set of the triplets from $\mathcal{K}_{\text{init}}$ as demonstrating examples for knowledge generation, and we repeat this T times for a given (image, question) pair (I, Q), then we can obtain T diversified knowledge statements for (I, Q). To further ensure that every time the demonstrating examples themselves are diverse, we first use K-means clustering to cluster the triplets in $\mathcal{K}_{\text{init}}$. Denote these K clusters as $\mathcal{K}_{\text{init}}^1, \mathcal{K}_{\text{init}}^2, \ldots, \mathcal{K}_{\text{init}}^K$. To generate T final knowledge statements for a given (I, Q) pair during the knowledge diversification step, we repeat the following process T times: (1) we randomly select one triplet from each $\mathcal{K}_{\text{init}}^k$, except the cluster the given (I, Q) pair belonging to, to form $K - 1$ demonstrating examples; (2) we use these $K - 1$ demonstrations as in-context examples to generate a knowledge statement for (I, Q), using the prompt template as described earlier. We call this strategy *self-supervised* knowledge diversification because we do not require any human to annotate diversified demonstrating examples. We will empirically compare this cluster-based strategy with a random demonstration selection strategy in our experiments. Details of how K-means clustering is done can be found in Appendix A.

3.2 Knowledge Integration for K-VQA

With the final set of T knowledge statements generated for each (image, question) pair, we can combine them with the image captions and the question, and pass them to a pre-trained text-based QA model for answer generation. In our experiments, we use UnifiedQA (Khashabi et al., 2020), OPT (Zhang et al., 2022) and GPT-3 (Brown et al., 2020).

4 Experiments

4.1 Datasets and Evaluation Metrics

To validate our proposed method, we choose two commonly used K-VQA benchmark datasets, namely, OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022). Questions in OK-VQA need outside knowledge beyond the images to answer. A-OKVQA is an augmented version of OK-VQA that requires additional types of world knowledge. Because the ground-truth answers of the *test-split* of A-OKVQA are not available, we use its *val-split* for evaluation. In the end, the OK-VQA and A-OKVQA datasets we use contain $5,046$ and $1,100$ questions, respectively. We report the soft accuracy (Goyal et al., 2017) on both datasets as there are multiple ground-truth answers for a question. Due to the limit of space, implementation details are provided in Appendix B.

4.2 Zero-shot Methods for Comparison

In this work, we focus on zero-shot K-VQA. There are models that need extra training (with labeled data other than K-VQA data). There are also some few-shot K-VQA methods where the few shots are dynamically selected from a large pool of training examples, which means they still need much training data. For fair comparison, we do not include these methods because they are not strictly zero-shot.

Below we briefly review three existing zero-shot K-VQA methods that we compare with:

PICa (Yang et al., 2022) converts images into captions with an off-the-shelf caption generator, CLIP-Cap (Mokady et al., 2021). The captions are regarded as contexts and fed to GPT-3 together with the question for answer prediction.

PNP-VQA (Tiong et al., 2022) uses improved caption generation by exploiting an image-text matching model (Li et al., 2022) to highlight image regions related to the question. The attended images are then used for caption generation with BLIP (Li et al., 2022) so that the captions are question-aware. We adopt the same caption generation method in PNP-VQA in our method. PNP-VQA uses UnifiedQA (Khashabi et al., 2020), a pre-trained question answering model, in a fusion-in-decoder (FiD) manner (Izacard and Grave, 2021), for final answer prediction.

Img2LLM (Guo et al., 2022) follows the caption generation process in PNP-VQA. Based on the captions, it generates synthetic QA pairs as demonstrating examples when prompting the LLM for final answers. OPT (Zhang et al., 2022) is used as the LLM for QA.

4.3 Main Results

In this section, we empirically evaluate our *generate and answer* approach in two ways: (1) We test the usefulness of the generated knowledge for K-VQA by systematically comparing our K-VQA system with and without knowledge generation. (2) We compare our *generate and answer* method with SOTA zero-shot K-VQA baselines, which do not explicitly generate knowledge.

The effect of knowledge generation. We first conduct systematic experiments to compare the *generate and answer* approach and the *directly answer* approach based on our own implementation. To see whether knowledge generation can consistently help K-VQA, we experiment with three different pre-trained QA models: UnifiedQA, OPT, and GPT-3. We choose these models because they are used in previous zero-shot K-VQA methods, namely, PNP-VQA, Img2LLM, and PICa, respectively. When using UnifiedQA, we follow Tiong et al. (2022) and adopt the FiD strategy. When using OPT, we follow Guo et al. (2022) and add synthetic QA pairs as demonstrations.[2]

Model, Size		Setting	OK-VQA	A-OKVQA
U.QA	0.7B	*w/o* KGen	32.3	29.0
		w KGen	39.7	31.6
	3B	*w/o* KGen	39.6	35.5
		w KGen	44.5	36.5
	11B	*w/o* KGen	43.7	38.9
		w KGen	45.4	39.1
OPT	6.7B	*w/o* KGen	35.2	32.4
		w KGen	39.2	35.9
	13B	*w/o* KGen	37.3	35.1
		w KGen	40.2	36.0
	30B	*w/o* KGen	37.7	34.4
		w KGen	42.2	38.1

Table 1: Performance comparison between using and not using generated knowledge. KGen refers to knowledge generation. **U.QA** is short for UnifiedQA.

LLM	Num. Kn.
w/o Gen. Kn.	39.6
LLaMA$_{7B}$	42.1
LLaMA$_{13B}$	42.5
GPT-3	44.5

Table 2: Results on OK-VQA when using generated knowledge from different models. *w/o Gen. Kn.* denotes without using any generate knowledge. The text-based QA model is UnifiedQA$_{3B}$.

We first show the results of UnifiedQA and OPT on both datasets in Table 1. We can see that under all settings (with different QA models and different model sizes), using the generated knowledge consistently improved the final accuracy of the answers. For GPT-3, due to the API cost, we only use the first 500 questions in OK-VQA for performance comparison. We find that on these 500 test examples, the answer accuracy increased from **27.4** to **34.1**, after adding generated knowledge.

Recently, a few open-source LLMs such as LLaMA (Touvron et al., 2023) have demonstrated

[2] We used the authors' code for synthetic QA pair generation. However, due to different implementation details and the different numbers of synthetic QA pairs used, the performance of our re-implemented Img2LLM base model differs from the reported performance.

Model	Accuracy
*Previous Zero-shot Models **without** Extra Training*	
PICa$_{\text{zero,175B}}$	17.7
PNP-VQA$_{0.7\text{B}}$	27.1
PNP-VQA$_{3\text{B}}$	34.1
PNP-VQA$_{11\text{B}}$	35.9
Img2LLM$_{6.7\text{B}}$	38.2
Img2LLM$_{13\text{B}}$	39.9
Img2LLM$_{30\text{B}}$	41.8
KGenVQA (Ours)	
UnifiedQA$_{3\text{B}}$	44.5
UnifiedQA$_{11\text{B}}$	**45.4**
OPT$_{30\text{B}}$	42.2
*Zero-shot Models **with** Extra Training*	
BLIP-2(OPT)$_{6.7\text{B}}$	36.4
BLIP-2(FlanT5$_{\text{XL}}$)$_{3\text{B}}$	40.7
BLIP-2(FlanT5$_{\text{XXL}}$)$_{11\text{B}}$	45.9
Flamingo$_{3\text{B}}$	41.2
Flamingo$_{9\text{B}}$	44.7
Few-shot Models (n=1)	
PICa$_{\text{few,175B}}$	40.8
PromptCap$_{175\text{B}}$	**48.7**

Table 3: Comparison with SOTA on OK-VQA.

comparable performance to GPT-3. We have also considered LLaMA as an alternative choice to GPT-3 for knowledge generation. We incorporate the generated knowledge into UnifiedQA$_{3\text{B}}$ for answer prediction. The results from using LLaMA generated knowledge are provided in Table 2. According to the results, we can conclude that incorporating generated knowledge from open-source LLMs also benefits K-VQA. By increasing the size of the LLMs, the generated knowledge can more effectively facilitate the model to arrive at the final prediction. In summary, the results demonstrate that the *generate and answer* approach consistently outperforms the *directly answer* approach on both benchmark datasets under different settings.

Although our main focus is the zero-shot setting, we also experiment with the few-shot setting, and we find that there is consistent improvement of the *generate and answer* approach over the *directly answer* approach in the few-shot setting, indicating the generalization of our method to few-shot settings. Details of our few-shot experiments can be found in Appendix C.

Comparison with SOTA. Next, we compare our method with the state-of-the-art models. Because we focus on zero-shot K-VQA without extra training, we only compare with previous models of this nature. The comparison is shown in the top half of Table 3 for OK-VQA and top half of Table 4 for A-OKVQA. We can observe the following from the tables: (1) On both datasets, our *KGenVQA* performs better than the zero-shot baselines when model sizes are comparable. For example, on OK-VQA, our UnifiedQA 3B surpasses all previous zero-shot baselines, i.e., baselines shown in the first block of Table 3. On A-OKVQA, our UnifiedQA 3B only loses out to Img2LLM 30B, but this is expected because of huge difference of model size. Our method with larger model sizes (i.e., our UnifiedQA 11B and OPT 30B) outperform all zero-shot baselines without extra training.

Model	Accuracy
*Zero-shot Models **without** Extra Training*	
Img2LLM$_{6.7\text{B}}$	32.3
Img2LLM$_{13\text{B}}$	33.3
Img2LLM$_{30\text{B}}$	36.9
KGenVQA (Ours)	
UnifiedQA$_{3\text{B}}$	36.5
UnifiedQA$_{11\text{B}}$	**39.1**
OPT$_{30\text{B}}$	38.1
Few-shot Models (n=10, 32 respectively)	
PICa$_{\text{few}}$	18.1
PromptCap$_{175\text{B}}$	**56.3**

Table 4: Comparison with SOTA on A-OKVQA.

We also show those zero-shot models with extra training (e.g., BLIP-2 (Li et al., 2023a), Flamingo (Alayrac et al., 2022)) and few-shot learning models (e.g., PICa$_{\text{few}}$ (Yang et al., 2022) and PromptCap (Hu et al., 2022)). It is worth noting that strictly speaking, PICa$_{\text{few}}$ (Yang et al., 2022) and PromptCap (Hu et al., 2022) do not use the same set of few shot examples (i.e., is not few-shot learning in the traditional sense) because these two methods dynamically sample demonstrating examples from the whole K-VQA training set for each test example. Because of their benefits from either extra training or access to the entire training set, we place these models in a different category, at the bottom half of Table 3 and Table 4. Compared with these models, we can see that our KGenVQA models still surpass some models with extra training, such as BLIP-2 (FlanT5$_{\text{XL}}$) and the powerful 3B Flamingo, and achieve comparable results with 9B Flamingo, demonstrating the effectiveness of our model compared with state-of-the-art models. Even comparing with few-shot models, we observe that our best performance is higher than PICa$_{\text{few}}$ (Yang et al., 2022) and is comparable to PromptCap$_{175\text{B}}$.

It may be worth noting that on OK-VQA,

Case	Num. Kn.	OK-VQA
Manual	1	35.9
Random	10	41.8
CoT	1	37.5
KGen	10	**44.8**

Table 5: Comparison of different knowledge generation methods on OK-VQA. "Num. Kn." is the number of knowledge statements used.

PICa$_{\text{zero}}$ performs poorly probably because it uses a single image caption. In order to make a fair comparison with PICa$_{\text{zero}}$, we provide results of our method with a single image caption and without image descriptions (i.e., with generated knowledge only) in Appendix D. The results show steady improvements (about $\mathbf{16}$ percentage points in terms of absolute accuracy) on OK-VQA.

4.4 Ablation Studies

Knowledge generation method. We first compare our cluster-based knowledge diversification strategy with (1) using the manual prompt generated knowledge, i.e., a single piece of knowledge (Manual); (2) randomly sampling $K-1$ single knowledge statement, instead of sampling from different clusters, from the initially generated knowledge statements, $\mathcal{K}_{\text{init}}$ for knowledge diversification in the second stage (Random). Besides, we consider the idea of Chain-of-Thoughts (CoT) (Wei et al., 2022), which generates explanations before the answer generation. In K-VQA, the needed knowledge can also be regarded as a kind of explanations. Therefore, we test the widely used CoT for knowledge generation, which is an alternative to our cluster-based knowledge generation approach. We re-use the six manual demonstrations as mentioned in Section 3 and manually add answers to the questions (i.e., each demonstration consists of contexts of image descriptions, a question, a piece of related knowledge and an answer). Together with these demonstrations, we prompt GPT-3 (Brown et al., 2020) to first generate the relevant knowledge and then the answer (CoT). Due to the cost of calling GPT APIs, we only apply CoT to a subset questions on OK-VQA (200 questions). We show model performance, based on UnifiedQA$_{3\text{B}}$, with different ways of knowledge generation and show results in Table 5. We have a few observations: (1) using initial generated knowledge with demonstrations offers improvements but no better than KGen. This may be that fixed manual demonstrations fail to generate diverse knowledge. For a fair comparison, we also consider using a single piece of knowledge from KGen, which achieves $\mathbf{38.8}$, indicating the need of diverse prompts in knowledge generation. (2) Comparing using random selection and cluster-based selection in the self-supervised knowledge diversification stage, we find that using the cluster-based method clearly outperforms random selection, which may not generate diverse knowledge. Overall, the cluster-based knowledge generation method is better than the other methods for knowledge generation in term of K-VQA performance; (3) When we compare the CoT knowledge generation with cluster-based knowledge generation, the second method significantly wins CoT in terms the benefit to K-VQA, probably because the cluster-based method has higher chances of facilitating answer generation with diverse knowledge; Besides, we also compare the direct CoT-generated answers from GPT-3 with answers generated when prompting GPT-3 for QA incorporating our generated knowledge. Our generated knowledge results in an accuracy of 32.0 while CoT-generated knowledge leads to 29.3.

QA Model	Num.	OK-VQA
UnifiedQA (FiD)$_{3\text{B}}$	0	39.6
	5	**44.5**
	10	**44.5**
	20	42.7
OPT$_{13\text{B}}$	0	37.3
	5	**40.2**
	10	37.2
	20	37.2
GPT-3	0	27.4
	5	**34.1**
	10	32.4
	20	31.7

Table 6: Performances with different numbers of knowledge statements.

Number of knowledge statements. Next, we test how the number of knowledge statements affects the performance, using UnifiedQA$_{3\text{B}}$ (FiD), OPT$_{13\text{B}}$ and GPT-3. Due to the API costs, we choose OK-VQA as the experiment dataset for this ablation study. For GPT-3 as the QA model, we test the performance on the first 500 questions. The results are reported in Table 6. Intuitively, we observe improvements after adding more generated knowledge at first and then decrement of performance. This is probably because adding too many pieces of knowledge may potentially add noisy or redundant

Case	Gram.	Rel.	Fact.	Help.
Ours$_{\text{max}}$	100.0	100.0	96.3	90.0
Ours$_{\text{avg}}$	99.0	100.0	94.5	67.0

Table 7: Evaluation of our generated knowledge in terms of four evaluation metrics.

knowledge, which harms the performance. Besides, we notice that decoder-only models have smaller optimal number of knowledge statements than encoder-decoder FiD model. This is probably because decoder-only models (i.e., OPT and GPT-3) may have difficulty in understanding the long concatenated sentence while FiD is specifically designed for comprehension of multiple documents.

4.5 Evaluation of the Generated Knowledge

In this section, we conduct human evaluation to exam the quality of the generated knowledge. We follow Liu et al. (2022) and sample 40 cases from OK-VQA dataset where the correctness of the answers would be changed (i.e., either from correct to wrong or wrong to correct) after adding the generated knowledge. For each instance, we sample 5 knowledge statements for evaluation. We ask two annotators to check the quality of the generated knowledge in terms of the evaluation metrics below. To ensure objectiveness, annotators will not know whether the predictions are changed to become correct or wrong.

Evaluation metrics. Following Liu et al. (2022); Shwartz et al. (2020), we take four metrics for evaluating generated knowledge: 1) *Grammatically*: whether it is grammatical 2) *Relevance*: whether it is related to answering the question and the image; 3) *Factuality*: whether it is factual; 4) *Helpfulness*: whether it is helpful so that it directly leads to the correct answers or provides indirect but supportive information of the correct answers. For *helpfulness*, we adopt three categories of evaluation: helpful (i.e., provides direct or indirect supportive information to correct answers), harmful (i.e., negates correct answers or support incorrect answers) or neutral (neither helpful or harmful). Besides the previously used metrics, we also consider *Diversity* as the fifth evaluation criteria, indicating the coverage of generated knowledge. Details about the definitions can be found in Appendix I and the examples we provide to annotators regarding the four evaluation metrics are included in the supplementary materials.

Results. The average agreement from two annotators over four evaluation metrics is 0.67, in terms of *Fleiss Kappa* κ (Landis and Koch, 1977). It indicates substantial agreement among annotators. For each criterion, we report the average score over two annotators. We consider two evaluation settings for generated knowledge: 1) *average*: taking the average scores over five pieces or knowledge; 2) *max*: take the maximum score over scores of five knowledge. The results are provided in Table 7. According to the results, most knowledge is grammatical, relevant to questions and factual. One interesting thing is that the generated knowledge may be relevant to questions but harmful for final answers, as the average score in term of *helpfulness* is only around 70. From the comparison with *average* and *max* scores of human evaluation, we further verify the need of knowledge diversification, which can raise the chance of generating helpful knowledge, as indicated by the maximum score of *helpfulness*, which means how likely the generated knowledge will lead to the correct answer. For diversity, we compare the five pieces knowledge generated by cluster-based selection against random selection. The average diversity of cluster-based select is $\mathbf{3.4}$, while $\mathbf{2.5}$ for random selection. It shows cluster-base selection results in more diverse knowledge, which is more likely to cover information for answering questions. It is in consistency with results in Table 5.

4.6 Case Study

To better understand the advantage of our method, we compare our method with the baseline, UnifiedQA$_{3B}$ (FiD), without generated knowledge. We analyze the first 20 cases, without cherry picking, where our method answers correctly while the baseline gives wrong predictions. Among the 20 error cases of the baseline, 85% are due to the lack of external knowledge, highlighting the advantage of our method. Due to the limitation of space, we provide the examples in Appendix G.

Besides, we conduct error analysis to better understand the limitations of our method. We conduct an empirical analysis for the error cases by manual checking 40 error cases from UnifiedQA$_{3B}$ (FiD) after adding generated knowledge. Among all error cases, we observe 20% are due to the undesired knowledge. Due to limitation of space, we provide visualization of the error cases in Appendix 4.6. The main cause of generating misleading knowledge comes from the inaccurate image descriptions which lack details for LLMs for knowledge generation. It implies with the development of better image description generation tools, our method can be potentially improved.

5 Conclusions

In this work, we propose to generate relevant knowledge from LLMs for zero-shot K-VQA. We evaluate the effectiveness of the generated knowledge by experimenting with different pre-trained QA models of varying model sizes on two K-VQA benchmarks. The experiment results show that the generated knowledge improves K-VQA performance, and our method can outperform SOTA zero-shot K-VQA methods. We further conduct human evaluation to validate the quality of the generated knowledge. The results demonstrate that the generated knowledge statements are relevant and helpful to questions in K-VQA.

6 Limitations

In this paper, we adopt GPT-3.5 as the LLM to generate several pieces of knowledge for one question. However, the generated knowledge may be redundant in some cases, which introduces noise to the final answer prediction process. Therefore, in the future, we need to investigate how to filter out redundant knowledge. Besides, in this work we only consider inserting the generated knowledge into a text-QA model when converting K-VQA into a text-based QA problem. A future direction is to design and insert generated knowledge into pre-trained vision-language models (PT-VLMs) (e.g., BLIP-2 (Li et al., 2023a)), because the conversion from images to texts may leave out crucial details, but PT-VLMs can take images as inputs without losing any potentially important visual information from the images.

Acknowledgement

This research was supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Grant No.: T2EP20222-0047, Project ID: MOE-000440-00). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

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, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 889–898.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *CoRR*, abs/2212.04037.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6325–6334.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and

Steven C. H. Hoi. 2022. From images to textual prompts: Zero-shot VQA with frozen large language models. *CoRR*, abs/2212.10846.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR*.

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

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

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP*, volume EMNLP 2020, pages 1896–1907.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1227–1235.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML*, volume 162, pages 12888–12900.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making large language models better reasoners with step-aware verifier. *CoRR*, abs/2206.02336.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 3154–3169.

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

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pages 2659–2670.

Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Computer Vision - ECCV*, pages 460–477.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV - 17th European Conference*, pages 146–162.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 4615–4629.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *CoRR*, abs/2210.01296.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C. H. Hoi. 2022. Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 951–967.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes.

In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 7371–7379.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2018. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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 *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 3081–3089.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Model and Size		# shots	Setting	OK-VQA
OPT	13B	32	*w/o* KGen	36.1
		32	*w* KGen	39.6
	30B	16	*w/o* KGen	36.7
		16	*w* KGen	43.8

Table 8: Performance comparison between using and not using generated knowledge in the few-shot setting on OK-VQA dataset. KGen refers to knowledge generation.

Model	Model Size
*Zero-shot Models **without** Extra Training*	
PICa$_{\text{zero}}$	175B
PNP-VQA	1.2B, 3.4B, 11.8B
Img2LLM	6.7B, 13B, 30B, 66B, 175B
Zero-shot Models with Extra Training	
VL-T5$_{\text{no-vqa}}$	269M
Frozen	7.1B
VLKD$_{\text{ViT-L/14}}$	832M
FewVLM	785M
BLIP-2(OPT$_{6.7\text{B}}$)	7.8B
BLIP-2(FlanT5$_{\text{XL}}$)	4.1B
BLIP-2(FlanT5$_{\text{XXL}}$)	12.1B
Flamingo	3B, 9B, 80B
Few-shot Models	
ClipCap→Cap.→GPT	175B
ClipCap→Ratl.→GPT	175B
PICa$_{\text{few}}$	175B
PromptCap	175B

Table 9: Summarizing of models for K-VQA.

A Details of K-Means Clustering

To divide testing instances into different clusters, we first convert each context-question-knowledge triplet into vector representations. Specifically, the context, question and the initial piece of knowledge will be concatenated and the textBERT (Devlin et al., 2019) to encode the concatenated sentence. Based on the encoded textual representation, we used the *K-Means* clustering to divide all instances into K clusters. Given an instance waiting for knowledge generation, which belongs to the cluster k, instances from other clusters will serve as demonstrations. In other words, we randomly select one demonstration from each cluster except the k-th cluster so that there are $K - 1$ demonstrations for the testing example. The set of demonstrations we denote as `PSEUDO DEMO`. Then we prompt LLMs again with the self-supervised demonstrations with an input. We will iteratively conduct the process mentioned above T times where at the t-th time step we obtain a piece of knowledge k_t and finally we have T knowledge pieces.

Model	Acc.
*Zero-shot Models **without** Extra Training*	
PICa$_{\text{zero},175B}$	17.7
PNP-VQA$_{0.7\text{B}}$	27.1
PNP-VQA$_{3\text{B}}$	34.1
PNP-VQA$_{11\text{B}}$	35.9
Img2LLM$_{6.7\text{B}}$	38.2
Img2LLM$_{13\text{B}}$	39.9
Img2LLM$_{30\text{B}}$	41.8
Img2LLM$_{66\text{B}}$	43.2
Img2LLM$_{175\text{B}}$	45.6
Zero-shot Models with Extra Training	
VL-T5$_{\text{no-vqa}}$	5.8
Frozen	5.9
VLKD$_{\text{ViT-L/14}}$	13.3
FewVLM	16.5
BLIP-2(OPT)$_{6.7\text{B}}$	36.4
BLIP-2(FlanT5$_{\text{XL}}$)$_{3\text{B}}$	40.7
BLIP-2(FlanT5$_{\text{XXL}}$)$_{3\text{B}}$	45.9
Flamingo$_{3\text{B}}$	41.2
Flamingo$_{9\text{B}}$	44.7
Flamingo$_{80\text{B}}$	50.6
Few-shot Models	
PICa$_{\text{few},175B}$ (n=1)	40.8
PromptCap$_{175B}$ (n=1)	48.7

Table 10: Model performancee on OK-VQA dataset. For models with different model sizes, we show the model size with subscripts.

B Experiment Settings

Experiment Details For knowledge generation, we use GPT-3.5 (*text-davinci-003*[3]) as our LLM, with a suggested temperature of 0.7. For the K-means clustering in knowledge diversification stage, we set the number of cluster to be 8 empirically.

For answer prediction, because exact match is adopted for evaluation, we encourage the pretrained QA model to give short answers. For UnifiedQA, we set the length penalty to be -1; for GPT-3.5, we add the following instruction: *Generate answers with as fewer words as possible.* After answer prediction, we conduct an answer post-processing step as proposed in (Awadalla et al., 2023).

We implement our model on NVIDIA Tesla V100 GPUs with 32 GB of dedicated memory. The system ran on CUDA version 11.1. For UnifiedQA, except 11B version, we implemented with a single GPU. For UnifiedQA 11B model and OPT model series, we implement with model parallel on four GPUs.

Package Version In this experiment, we rely on the PyTorch library, 1.13.1 version. For the implemen-

[3] https://platform.openai.com/docs/models/gpt-3-5

Model	Acc.
*Zero-shot Models **without** Extra Training*	
Img2LLM$_{6.7B}$	33.3
Img2LLM$_{13B}$	33.3
Img2LLM$_{30B}$	36.9
Img2LLM$_{66B}$	38.7
Img2LLM$_{175B}$	42.9
Few-shot Models	
ClipCap→Cap→GPT$_{175B}$ (n=10)	16.6
ClipCap→Rel→GPT$_{175B}$	18.1
PromptCap$_{175B}$ (n=32)	56.3

Table 11: Model performancee on A-OKVQA dataset. For models with different model sizes, we show the model size with subscripts.

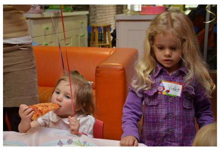
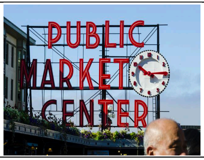
Img.		
Ques.	Which type of leather is used for making the sofa set shown in this picture?	Where in the world is this located?
GT.	cow, fake, fine grain, suede	seattle, san francisco, seattle usa, boston massachusetts
Pred.	black leather	czech republic
Cap.	two child a pizza pizza three people child up pizza. a young girl and a young girl with pizza as food. a young girl eating pizza while sitting in a booth	a sign outside of a market market sign on a clear day. the sign shows market square, with a lot of people, and a large clock. a group of people outside of a building showing a clock.
Kn.	The sofa set shown in this picture is likely made of faux leather, which is a synthetic material made to look and feel like real leather.	This market square is located in the city of Prague, Czech Republic.

Table 12: Visualization of error cases. GT. is for ground-truth annotation, Pred. is for predictions from models, Cap. is for the image captions and Kn. is for generated knowledge.

tation of BLIP (Li et al., 2022) (used for image caption generation), we leverage the LAVIS package from Salesforce [4] (version 1.0.2), for OPT (Zhang et al., 2022) and UnifiedQA model (Khashabi et al., 2020) we use the transformers package from Huggingface [5] (version 4.29.2), and for GPT-3.5 model, we leverage the OpenAI API [6].

Model Size: We show model size in Table 9. If we one model has different versions of model size, we separate them with comma.

C Few-shot Setting Results

We provide the results for our method in the few-shot setting on OK-VQA in the section. Specifically, we leverage the OPT model (Zhang et al., 2022) as the final QA model and give a few demonstrations. Each demonstration consists of a question, an image description as the context, an answer and optional related knowledge (in the *w* KGen setting). The results are shown in Table 8. According to the results, we observe consistent improvements after adding generated knowledge, indicating our method can generalize to the few-shot setting as well.

D Fair Comparison with PICa$_{zero,175B}$

Considering PICa$_{zero,175B}$ leverages only a single piece of image description while our method uses multiple captions, following (Tiong et al., 2022), improvements may potentially come from more detailed image descriptions. To ablate the impact from image description side, we use a single caption as the image description, similar to PICa$_{zero,175B}$. It achieves **33.8** on OK-VQA, with about **16** absolute accuracy improvements over PICa$_{zero,175B}$. Further more, we used only the generated knowledge as inputs to text-based QA models (UnifiedQA$_{3B}$). It achieves **33.5** on OK-VQA, highlighting that generated knowledge itself contains information for question answering.

E Model Performance

We only provide models in a fair comparison in Section 4.3. In this part, we provide performance of models on K-VQA including zero-shot K-VQA models without extra training but have larger model sizes, zero-shot K-VQA models with extra training and few-shot K-VQA models. The results on OKVQA and A-OKVQA are shown in Table 10 and Table 11 respectively.

F Error Cases

In this section, we provide visualization of two error cases of which the generated knowledge is inadequate. The reason of generating the harmful knowledge is because of inaccurate image captions. A potential way of improving our method is to improve the quality of image descriptions.

[4] https://github.com/salesforce/LAVIS/tree/main/lavis
[5] https://huggingface.co/
[6] https://platform.openai.com/overview

Image			
Question	What would happen if these items fall to the ground?	What sates are these grown in?	Name one famous person whom also has a black and white one of these?
Ground Truth	shatter, they would shatter, break, they would break	florida california, california, florida	taylor swift, russell brand, hillary clinton, ernest hemingway
Base Prediction	nothing	texas	kate winslet
Generated Knowledge	If a glass item falls to the floor, it will break.	California and Florida are the leading producers of oranges.	Taylor Swift is a famous singer and songwriter who has a black and white cat named Meredith.
Our Prediction	they would break	california	taylor swift
Image			
Question	If it gets cold enough what will happen to the area being stepped over?	What knocked the guy off his chair?	What is the white cloud behind the jet called?
Ground Truth	freeze, frozen, it will freeze over, iced	wave, water	contrail, cloud, supersonic wave
Base Prediction	snow	water splash	halo
Generated Knowledge	If it gets cold enough, the area being stepped over will freeze, creating a layer of ice on top of the snow.	The waves in the water knocked the man off his chair.	The condensation trail, or contrail, is a visible trail of condensed water vapor created by an aircraft engine or wingtip vortices under certain atmospheric conditions.
Our Prediction	frozen	wave	contrail

Table 13: Visualization of error cases of the baseline without generated knowledge, while our method answers correctly with the help with generated knowledge. Wrong predictions are highlighted in red.

G Comparison with the Baseline without Knowledge

In this section, we provide visualization of error cases of the baseline model without knowledge and compare with our method. The visualized examples are shown in Table 13. Noted, we do not perform cherry-picking. The visualized cases are the first six error cases of the baseline model on OK-VQA while being correctly addressed by our method. To keep the table tidy, we only present one piece of generated table. According to the visualization, we observe our generated knowledge largely benefit addressing these questions in need of external knowledge.

H Manual Prompts

Here we provide a full list of six manual prompts in Table 14. Before the demonstrations, we also add an instruction: *Please generate related background knowledge to the question:* in the front. Knowledge are collected from searching with Google.

I Details for Human Evaluation

In this part, we provide more details about human evaluation about the knowledge quality. We invite two annotators for evaluation of 40 questions with five pieces of generated knowledge. Firstly, they will be given an instruction, indecating the definition of the K-VQA task, an example of the K-VQA task and the goal of the evaluation. Next, we describe what information (i.e., question, ground-truth answer, generated knowledge, and image) will be provided to them and the denotations of the information. Thirdly, we elaborate the definitions of four metrics. For the metrics of *Relevance*, *Factuality* and *Helpfulness*, besides definitions, we provide a few concrete examples in texts to make it easier for understanding. The definifions and examples are provided in Table 15. For the full information of the annotated knowledge, please refer to the Supplementary file.

Num.	Content
1	Context:The company in the image is Monsanto. There are two men selling products. The logo behind two men is Monsanto. Question:What does company in the image own? Knowledge:Monsanto is a multinational agrochemical and agricultural biotechnology corporation. It is one of the world's leading producers of roundup, a glyphosate herbicide.
2	Context:The red vegetable is tomato. There is a sandwich with tomato and lettuce. There is a sandwich on the table. Question:Where can this red vegetable be found? Knowledge:tomatoes are usually planted in gardens.
3	Context:The man is playing tennis. The man is holding a tennis racket. A man is in a competition of tennis. Question:What English city is famous for a tournament for the sport this man is playing? Knowledge:The Wimbledon Championships is the oldest tennis tournament in the world.
4	Context:a plate with ham, tomatoes, meat, and sliced peppers on top of it. breakfast and bacon eggs scrambled toast. a breakfast sandwich, tomatoes, bacon, and eggs Question:what food in the photo has a lot of c vitamin? Knowledge:Tomatoes and tomato products are rich sources of folate, vitamin C, and potassium. Eggs contain decent amounts of vitamin D, vitamin E, vitamin B6, calcium and zinc. Bacon provides a good amount of B vitamins.
5	Context:a man sitting in front of a laptop computer smiling and posing for the camera. a man wearing glasses sitting in front of a laptop. a man in glasses and glasses at a desk with laptop. Question:what purpose do the glasses the man is wearing serve? Knowledge:Glasses are typically used for vision correction, such as with reading glasses and glasses used for nearsightedness.
6	Context:a bedroom with a bed, wall paper and lamp. a bed with storage underneath it in a room. a bed in a small room with pillows and box drawers. Question:what was the largest size of that platform that we have? Knowledge:Single size is 91 cm x 190 cm. Super single size is 107 cm x 190 cm. Queen size is 152 cm x 190 cm. King size is 182 cm x 190 cm.

Table 14: Contents of manual prompts.

Attributes	Definition	Example
Grammaticality	Whether the knowledge statement is grammatical (e.g., whether a complete and fluent sentence; whether human can understand the sentence).	None
Relevance	Whether a knowledge statement is relevant to the given question. A statement is relevant if it covers the same topic as the question or contains a salient concept that is the same as or similar to the one in the question (provided indirect but related information).	[Image]: a bedroom with a bed [Question]: what was the largest size of that platform that we have? [Knowledge]: Single size is 91 cm x 190 cm. Super single size is 107 cm x 190 cm. Queen size is 152 cm x 190 cm. King size is 182 cm x 190 cm. [Judge]:Relevant. Because the information is related to the topic on bed size.
Factuality	Whether a knowledge statement is (mostly) factually correct or not. If there are exceptions or corner cases, it can still be considered factual if they are rare or unlikely.	[Image]: a triangle in the image [Question]: what shape is the object in the image? [Knowledge]: A rectangle is a shape with two equal sides [Judge]: Not factual, because a rectangle has four sides [Image]: a limousine; a car [Question]: how many doors does the vehicle in the image have? [Knowledge]: A limousine has four doors. [Judge]: Factual. [Image]: a human being [Question]: how many fingers does this creature have? [Knowledge]: A human hand has four fingers and a thumb. [Judge]: Factual, despite that there are exceptions – people with disabilities may have less or more fingers.
Helpfulness	Whether a knowledge statement is (mostly) factually correct or not. If there are exceptions or corner cases, it can still be considered factual if they are rare or unlikely.	[Image]: a subway in the image [Question]: How often you take this transportation back and forth to work per week? [Knowledge]: You take the subway back and forth to work five days a week [Judge]: Helpful. Because the statement directly supports the answer. [Image]: a spider [Question]: how many legs does the animal in the image have? [Knowledge]: Arachnids have eight legs [Judge]: Helpful. Although the statement does not directly refer to spiders, together with the fact that "spiders are a kind of arachnids" it completes a reasoning chain in deriving the answer. [Image]: two persons are playing chess [Question]: what are the results of the game? [Knowledge]: A game of chess has two outcomes [Judge]: Harmful. Since the statement supports answering "two outcomes" instead of "three outcomes". [Image]: a person in the white background. [Question]: How many chromosomes does the creature have? [Knowledge]: human beings are mammals. [Judge]: Neutral. The knowledge does not provide information in favor or contrast of answering the question.

Table 15: Definitions and examples for evaluation metrics.