Learning to Ask Denotative and Connotative Questions for Knowledge-based VQA

Xiaoying Xing **Peixi Xiong** **Lei Fan** **Yunxuan Li** **Ying Wu**
Northwestern University
{xiaoyingxing2026, peixixiong, leifan, yunxuanli2019}@u.northwestern.edu
yingwu@northwestern.edu

Abstract

Large language models (LLMs) have attracted increasing attention due to its prominent performance on various tasks. Recent works seek to leverage LLMs on knowledge-based visual question answering (VQA) tasks which require common sense knowledge to answer the question about an image, since LLMs have obtained rich knowledge from large-scale training. Several methods have proposed to leverage frozen LLMs by converting visual information to textual prompts. However, how to efficiently exploit the knowledge of LLMs and bridge the disconnects between visual information and language models remain open problems. In this paper, we propose to let LLMs learn to ask (L2A) informative questions to collect essential visual information. We introduce the concepts of denotation and connotation to promote image and question understanding and provide a clear guidance with respect to the objective of question generation. In this way, the model can better capture the associations between different concepts, as well as efficiently collect both explicit information and implicit relevant information that contribute to the final answer. The experiments demonstrate that our proposed method achieves consistent performance on various knowledge-based VQA datasets.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023) have shown remarkable potential on various natural language tasks due to their rich knowledge and reasoning abilities obtained by large-scale pretraining. Recent works have also explored to leverage LLMs on vision-language tasks, especially knowledge-intensive tasks such as knowledge-based Visual Question Answering (VQA) (Marino et al., 2019; Schwenk et al., 2022), which aims to answer questions about images that require commonsense knowledge to answer.

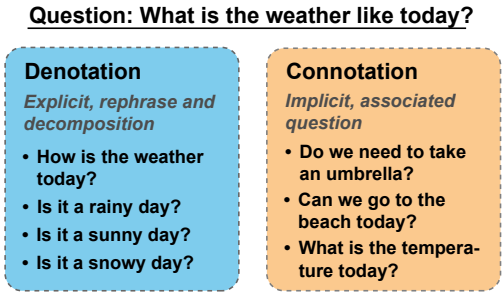

Figure 1: An intuitive illustration of sentence-level denotation and connotation with a simple example. Denotation refers to rephrasing or decomposition of the original question. Connotation refers to being implicitly associated and facilitate question understanding.

The main challenge of leveraging LLMs on vision-language tasks is to bridge the disconnect between visual information and language models. A straightforward solution is to train the models on large-scale multimodal data (Li et al., 2023; Alayrac et al., 2022), while is limited by expensive computation costs. Another line of works seek to use frozen LLMs as implicit knowledge source and convert the image contents to textual prompts to the LLMs. They mainly describe the visual contents by image captions (Shao et al., 2023; Yang et al., 2022; Hu et al., 2023; Tiong et al., 2022; Du et al., 2023; Chen et al., 2023b) and question-answer pairs about the image (Guo et al., 2023; Wang et al., 2023; Lan et al., 2023). However, the caption-based methods may drop essential visual information required to answer the question, since captions often provide a general description of the whole image. Although several methods (Hu et al., 2023; Tiong et al., 2022; Du et al., 2023) have proposed to improve the quality of the generated captions, encapsulating all the required information within a single description still remains challenging. On the other hand, the question-based methods generate questions about the images and can provide the LLMs with more exhaustive visual information

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8301–8315

through multiple question-answer pairs. Nevertheless, the generated questions may be irrelevant to the target question and introduce noisy information. Existing question generation methods lack a well-defined objective towards relevant and informative questions that are essential to the target task.

In this paper, we inspire LLMs to learn to ask (L2A) relevant questions with clear objectives. In order to provide guidance regarding informative question generation, we introduce the concepts of denotation and connotation that were originally proposed in linguistics studies (Sonesson, 1998; Rao, 2017). The original definition of denotation is the precise literal meaning of a word, while connotation primarily refers to the wide array of associations surrounding the word. Both of the concepts help with understanding the meaning of the words. We extend these concepts from word-level to sentence-level and prompt the LLM to generate both connotative and denotation questions with respect to the original question. In this way, the model can extract both explicit information and implicitly associated information to gain a more comprehensive understanding of the image contents as well as the question. An intuitive illustration of the extended concepts with a simple example are shown in Figure 1. Our proposed question generation strategy incorporates the exterior and inherent correlations between concepts, hence takes better advantage of the intrinsic knowledge of LLMs to efficiently extract relevant visual information.

We mainly evaluate our proposed method on OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) datasets, where both datasets ask questions that require open-world knowledge beyond the image to obtain the correct answers. The experiment results show that our proposed method achieves consistent performance without requiring in-context examples, and only requires a small number of question-answer pairs as prompting exemplars. Our main contributions are summarized as follows:

- We elaborately guide the LLMs to learn to ask informative questions with clear objectives, which enables them to directly generate high-quality relevant questions.
- We introduce the concepts of denotation and connotation and extend them to sentence-level question generation objectives, which inspires the model to extract both explicit and implicit information from images.
- Our proposed prompting strategy effectively harnesses the intrinsic knowledge of LLM and guides the LLM to fulfill the task without requiring in-context examples.

2 Related Works

2.1 Using LLMs for knowledge-based VQA

Knowledge-based VQA (Marino et al., 2019; Schwenk et al., 2022) requires external open-knowledge to answer the question about an image. Traditional methods (Lin et al., 2022; Wu et al., 2022) explicitly retrieve relevant knowledge from external knowledge resources. Recent works attempt to use LLMs as implicit knowledge source. They often convert image contents to textual prompts to LLMs and utilize the LLMs to integrate the information and reason about the final answer. A category of works describe the visual contents by image captioning (Yang et al., 2022; Hu et al., 2023; Tiong et al., 2022; Du et al., 2023) and object descriptions (Chen et al., 2023b). Another line of works represent the visual information by multiple question-answer pairs (Wang et al., 2023; Lan et al., 2023; Guo et al., 2023). Some works also incorporate the answer heuristics from different vision-language models (Lan et al., 2023; Shao et al., 2023). Despite the success of previous works in leveraging LLMs for knowledge-based VQA, it still remains an open problem to obtain a comprehensive description of the required visual information without introducing irrelevant contents.

2.2 Visual question generation

Visual question generation (VQG) (Mostafazadeh et al., 2016; Patro et al., 2018) task aims to generate questions related to a given image, which is often employed to improve the VQA performance. Early works mainly generate questions given specific answers (Li et al., 2018; Krishna et al., 2019; Liu et al., 2018) or answer categories (Uppal et al., 2021), and others (Shah et al., 2019; Lan et al., 2023) generate rephrasings of the questions to improve model robustness. Recent work Img2LLM (Guo et al., 2023) utilizes an extra model to generate questions conditioned on specific answers to prompt the LLM, while they generate large amount of questions and may introduce irrelevant information. FIIG (Wang et al., 2023) takes a further step and proposes a refinement model to filter the relevant questions, while still lacks a well-defined objective towards gener-

ating relevant questions. In contrast to previous methods, we propose a clear guidance to enable the LLM learn to directly generate informative questions. We take better advantage of the intrinsic knowledge of LLMs to actively collect essential information for answering the target question.

2.3 Prompt learning

Prompt learning aims to automatically generate appropriate prompts that effectively guide LLMs to produce desired outputs without extensive model training. It is widely discussed to refrain from the time-consuming process of prompt engineering and promote the performance of LLMs meanwhile. Prior works have developed soft prompt-tuning methods that learns continuous vectors as prompt (Qin and Eisner, 2021; Li and Liang, 2021; Lester et al., 2021). However, soft prompts have limited interpretability and generalizability across different models due to disparities in latent spaces. Another line of works alternatively learn discrete prompt optimization by gradient-guided search (Shin et al., 2020; Wen et al., 2023; Chen et al., 2023a) and reinforcement learning (Zhang et al., 2022; Deng et al., 2022; Diao et al., 2022). Recent works (Yang et al., 2023; Zhou et al., 2022; Pryzant et al., 2023) propose to leverage LLMs for prompt optimization. They show that LLMs are capable of generating and updating textual prompts based on feedback. In this paper we elaborately introduce prior knowledge to simplify the prompt learning procedure and provide effective guidance to the LLM by combining predefined static prompts and automatically generated question-specific prompts.

3 Method

Our proposed L2A approach motivates LLMs to learn to ask informative questions that collect essential information for answering visual questions. An overview of our proposed method is shown in Figure 2. It consists of three steps: prompt generation, information collection and information integration, which mainly involves the interaction between a inquirer (*i.e.,* LLM) and a respondent (*i.e.,* base VQA model). The prompt generation module aims to provide clear guidance for the inquirer to ask informative questions, by introducing the clarified concepts of denotation and connotation. During the information collection stage, under the guidance of the prompt, the inquirer generates both denotative and connotative questions to collect essential information. The respondent generates the corresponding answers based on the visual contents. Then we evaluate the obtained question-answer pairs by their contribution to the final answer, which implies their relevance to the target question. Finally, we leverage the implicit knowledge and reasoning ability of LLM to integrate the collected information, hence obtain more reliable answers to the knowledge-intensive questions.

3.1 Prompt generation

We first introduce the concepts of denotation and connotation. The original definition of denotation is the precise literal meaning of a word (*e.g.*, the denotative meaning of *home* is *a dwelling place*). Connotation originally refers to the wide array of associations surrounding the word (*e.g.*, the connotative meaning of *home* is *comfort, love, security or privacy*). Both of them help with understanding the meaning of the words, while denotation enhances the specific interpretation and connotation extends the understanding by associating with other concepts. To guide the objective of question generation using these concepts, we extend the word-level definition of denotation and connotation to sentence-level, as illustrated in Figure 1. We define denotative and connotative questions of the target question as follows:

- *Denotative questions are either the rephrasing or decomposition of the original question, which extract explicit information.*
- *Connotative questions seek implicitly associated information that can help in understanding and answering the original question.*

Then we propose a prompting template combining the global guidance of denotative and connotative questions with question-specific inspiring prompt. The template is as follows:

```
Given a question, define its denotative ques-
tions and connotative questions.
/* Define denotation & connotation */
Denotative questions are either ...
Connotative questions seek implicitly ...
-----------------------------------
Please generate denotative and connotative
questions of the target question.
/* Question-specific input */
Caption: <c> Question: <q>
Inspiring words: <w1, w2, ···>
```

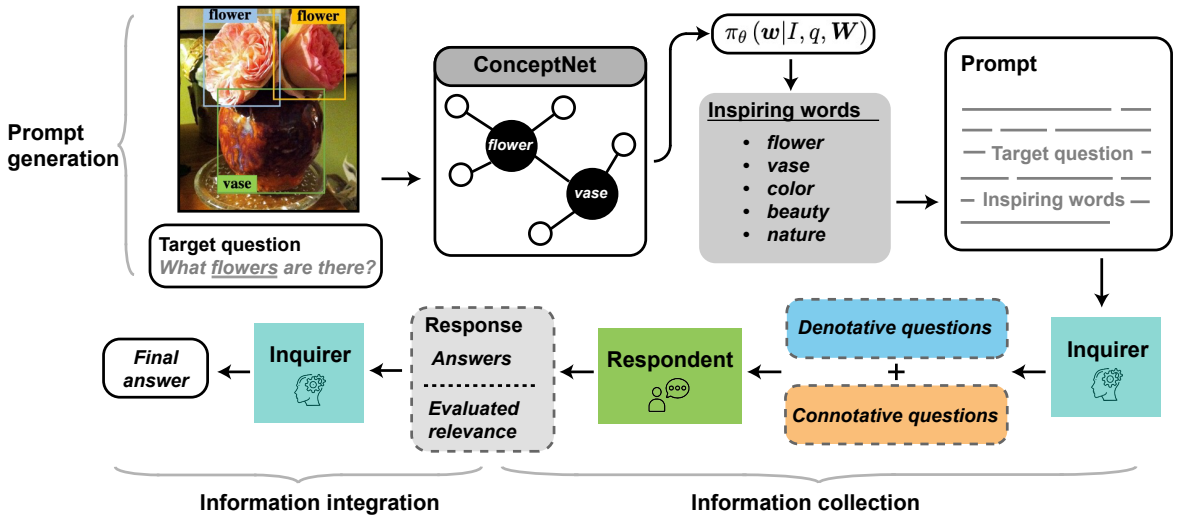

Figure 2: Illustration of the inference process of the proposed L2A approach. It consists of Prompt generation, Information collect and Information integration modules. Guided by the generated prompt, the inquirer generates denotative and connotative questions to collect visual information. Then the respondent gives back the answers to the generated questions and the evaluated relevance of them. Finally the inquirer integrates the collected information to decide the final answer to the target question.

The inspiring words are learned prompting words based on the input question and image. They are introduced to (1) incorporate the prior knowledge of conceptual correlations and introduce the concepts related to the question and image. (2) bridge the gap between the original word-level definition of denotation and connotation with the sentence-level generation objective. (3) allow for the optimization of the prompt. Specifically, given input image I and question q, we first extract the keywords from the q by part-of-speech parsing (Honnibal and Montani, 2017), as well as the object tags from the image by off-the-shelf object detection model (Redmon et al., 2016). Then we select the words that are connected with them in ConceptNet (Speer et al., 2017) as candidates, denoted as $\boldsymbol{W}$. We aim to learn a policy to select a series of words $\boldsymbol{w} = [w_1, \cdots, w_T]$ from $\boldsymbol{W}$. At each time step t, the model generates the next word w_t conditioned on previous tokens $\boldsymbol{w}_{<t}$. We implement the policy network with a light-weight sequence-to-sequence model following (Zhang and Zhu, 2021). Suppose the policy network is parameterized with θ, the problem can be formulated as:

$$\max_\theta R(\boldsymbol{w}, I, q), \ \boldsymbol{w} \sim \prod_{t=1}^{T} \pi_\theta\left(w_t|I, q, \boldsymbol{W}, \boldsymbol{w}_{<t}\right) \tag{1}$$

The reward R is the accuracy of target question answering given the generated questions and corresponding answers as prompts, and the policy is updated by REINFORCE (Williams, 1992) algorithm. The generated inspiring words $\boldsymbol{w}$ fill in the prompt template to constitute the full prompt $p_{\boldsymbol{w}}$.

3.2 Information collection

In the information collection stage, the inquirer agent is prompted to ask denotative and connotative questions guided by $p_{\boldsymbol{w}}$. An initial caption c is also provided to generate questions related to the image. Denoting the series of generated questions as $\boldsymbol{q'}$, the k-th question among $\boldsymbol{q'}$ is generated as:

$$q'_k = \arg\max_{\tilde{q}'_k} P_{\text{LLM}}\left(\tilde{q}'_k|q, c, p_{\boldsymbol{w}}\right) \tag{2}$$

Then it receives the answer from the respondent agent, denoted as a'_k. In order to provide more information for the inquirer to reason about the answer to the target question, we evaluate the $q'_k a'_k$ pairs by their contribution to the final answer. Previous metrics to evaluate question generation often measure the similarity between the generated questions and the target question, while exterior textual similarity differs from interior semantic relevance in many cases (Zhong et al., 2008). In contrast, we propose to evaluate the question relevance in a task-specific manner by their contribution to the final answer. We concatenate each generated question-answer pair as $[q'_k; a'_k]$, which can be regarded as a fact f_k that provides evidence to the final answer. The contribution of f_k can be represented as the information gain it brings, denoted as G_k. Suppose

the each answer candidate to q is a_i, then G_k can be formulated as:

$$G_k = \sum_{a_i} -P(a_i|I,q)\log P(a_i|I,q) - \sum_{a_i} -P(a_i|I,q,f_k)\log P(a_i|I,q,f_k) \quad (3)$$

The corresponding scores are used to estimate the relevance of the generated question-answer pairs to the target question.

3.3 Information integration

During information integration stage, the inquirer summarizes the collected information to decide the final answer $\hat{a}$ to the target question, which can be represented as:

$$\hat{a} = \arg\max_{\tilde{a}} P_{\text{LLM}}\left(\tilde{a}|q,c,\boldsymbol{q'},\boldsymbol{a'},\boldsymbol{G}\right) \quad (4)$$

where the relevance score G_k serves as implicit weighting factor upon the generated question q'_k and the answer a'_k. In this way, the inquirer can leverage its internal knowledge and reasoning abilities to alleviate the impact of irrelevant questions. The whole instruction prompt is as follows:

```
/* Answer instruction */
Given some related question-answer pairs and their relevance to the target question, integrate the information and give the short answer to the target question.
Caption: <c> Target question: <q>
/* Collected information */
Denotative questions:
Question: <q'_1> Answer: <a'_1> Relevance: <G_1>
...
Connotative questions:
Question: <q'_i> Answer: <a'_i> Relevance: <G_i>
...
Answer:
```

4 Experiments

4.1 Experiment settings

Datasets. We mainly evaluate our proposed method on knowledge-based VQA datasets OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022). Both of the datasets ask questions that require open-world knowledge beyond the image to answer. OK-VQA dataset contains 14,055 image-question pairs associated with 14,031 images from MSCOCO dataset (Lin et al., 2014). Each question is annotated with ten open-ended answers. A-OKVQA is an augmented benchmark containing 25K image-question pairs. It encompasses both multiple-choice (MC) settings and direct-answer (DA) settings without answer options. Both datasets employ the soft accuracy (Antol et al., 2015) as the evaluation metric.

Implementation details. We implement the inquirer agent by gpt-3.5-turbo[1]. We employ BLIP-2 FlanT5$_{\text{XL}}$ (Li et al., 2023) as the respondent and provide initial image captions to facilitate the question generation. The number of questions to ask is decided dynamically by the inquirer according to the demand of information collection, and the average number for each target question is 8.5.

Compared methods. We mainly compare with the methods that also leverage large-scale language models to answer the questions. We categorize the compared methods by how they utilize LLM to answer the questions for fair comparison as (1)methods directly trained on large-scale multimodal data, (2)methods prompting frozen LLMs to answer the questions with a few in-context examples, (3)methods prompting frozen LLMs without examples. Among the methods that prompt frozen LLMs for question answering, Img2LLM (Guo et al., 2023) and FIIG (Wang et al., 2023) also involve converting visual contents into question-answer pairs. Img2LLM utilizes extra question-generation model to generate questions given pre-extracted answers. They generate 30 question-answer pairs and 100 image captions to answer each question, which may introduce redundant information. FIIG first generates general questions then filters the relevant questions, and requires multiple examples to guide the LLM for question generation and answering.

4.2 Main results

Comparison with state-of-the-arts. The results on OK-VQA test set and A-OKVQA validation set are shown in Table 1. The column *Shot number* refers to the number of in-context examples to prompt the LLM and *Exemplar number* represents the total number of exemplars (*e.g.*, captions, answer heuristics or question-answer pairs) provided to the LLM. Our proposed method achieves consistent performance on both datasets among the methods that prompt LLM without in-context examples or directly train on multimodal data. We do not require large computation costs or extra anno-

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

Method	Shot number	Exemplar number	OK-VQA	A-OKVQA	
				DA	MC
Models directly trained on multi-modal data					
FewVLM$_{\text{large}}$ (Tan and Bansal, 2019)	0	0	16.5	-	-
Flamingo$_{\text{80B}}$ (Alayrac et al., 2022)	0	0	50.6	-	-
BLIP-2 FlanT5$^{\dagger}_{\text{XL}}$ (Li et al., 2023)	0	0	40.7	34.6	47.5
BLIP-2 FlanT5$^{\dagger}_{\text{XXL}}$ (Li et al., 2023)	0	0	45.9	37.4	50.4
Using frozen LLM for answering w/ in-context examples					
PICa-Full (Yang et al., 2022)	16	16	48.0	-	-
Prophet (Shao et al., 2023)	20	20	61.1	58.2	76.4
Prophet+FIIG (Wang et al., 2023)	20	60	61.3	59.8	-
PromptCap (Hu et al., 2023)	32	32	60.4	56.3	73.2
IPVR (Chen et al., 2023b)	8	8	44.6	46.4	-
TOA (Xing et al., 2023)	16	16	60.6	61.2	63.1
Using frozen LLM for answering w/o in-context examples					
PNP-VQA$^{\dagger}_{\text{3B}}$ (Tiong et al., 2022)	0	100	34.1	35.1	53.1
PNP-VQA$^{\dagger}_{\text{11B}}$ (Tiong et al., 2022)	0	100	35.9	36.3	53.5
Img2LLM$_{\text{175B}}$ (Guo et al., 2023)	0	30	45.6	42.9	-
Img2LLM+RQP (Lan et al., 2023)	0	30	46.4	43.2	-
LAMOC$_{\text{11B}}$ (Du et al., 2023)	0	10	40.3	37.9	-
L2A (ours)	0	8.5	46.2	48.5	62.4

Table 1: Comparison to the state-of-the-art methods on knowledge-based VQA datasets. † denotes some of the results are based on reimplementation. DA refers to direct-answer setting and MC refers to multiple-choice setting. *Shot number* refers to the number of in-context examples. *Exemplar number* represents the total number of exemplars (*e.g.*, captions, answer heuristics or question-answer pairs) provided to the LLM.

tations of examples to prompt the LLMs, and we only prompt the LLM with a small number of exemplars. It indicates that our prompting strategies effectively guide the LLM to fulfill the required task. With the clear objective to guide the question generation, the LLM can learn to directly ask essential questions that efficiently collect relevant visual information. Besides, we surpass the base VQA model BLIP-2 FlanT5$_{\text{XL}}$, which is used to answer the generated questions in our method, by a large margin. It further demonstrates that by generating multiple denotative and connotative questions, the model attains more comprehensive image understanding and obtains additional visual information to enhance the question answering.

Answer analysis. Recent works using LLMs to directly answer open-ended questions may generate answers that are semantically equivalent to the ground truth but use different expressions (*e.g.,* pizzeria vs. pizza restaurant, herbivorous vs. eat grass, cycling vs. riding bicycle). The conventional exact matching evaluation is not applicable in many cases and often underestimate the model capability. Therefore, we conduct complementary experiments to evaluate the answers using an independent LLM. Specifically, for the answers unable to exactly match with the ground truth, we use the LLM to discern the equivalent answers in context of the question similar to (Mañas et al., 2023). We compare the results under different evaluation methods for indication in Table 2. It should be noted that some methods in Table 1 have implicitly involved knowledge of answer vocabulary in the training process or prompting examples, hence their performance may not be underestimated by the evaluation metric. To better reflect the model's real capabilities, we adopt the aforementioned evaluation method in subsequent experiments.

4.3 Ablation Studies

To verify the effectiveness of each component of our proposed method, we conduct several ablation studies on OK-VQA dataset, as shown in Table 3.

Prompting strategy The main contribution of our method is to define the objective of informative question generation and propose an effective prompting strategy. Therefore, we conduct ablative experiments that (1) remove the instructions

Method	OK-VQA		A-OKVQA	
	Match	LLM	Match	LLM
BLIP-2$_{XL}$	40.7	48.2(↑ 7.5)	34.6	40.7(↑ 6.1)
PNP$_{3B}$	34.1	37.6(↑ 3.5)	35.1	40.2 (↑ 5.1)
PNP$_{11B}$	35.9	39.6(↑ 3.7)	36.3	43.7(↑ 7.4)
L2A	46.2	54.1(↑ 7.9)	48.5	55.6(↑ 7.1)

Table 2: Comparison of results using conventional exact-match evaluation and LLM based evaluation.

Model	Accuracy
L2A-full	54.1
w/o instruction	50.8
random	53.7
w/o evaluation	53.7

Table 3: Ablation studies on important components of the proposed method.

Generation method	Accuracy
slot-filling	42.7
free-form	44.6
LLM	50.4
L2A	53.9

Table 4: Comparison of different question generation strategies.

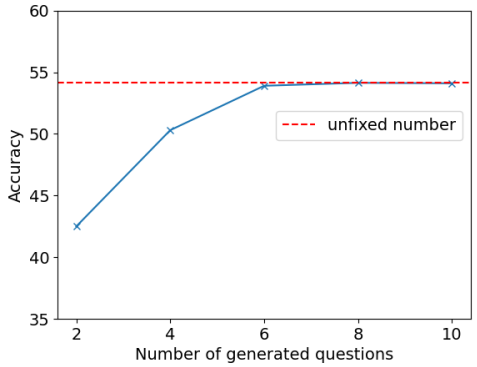

Figure 3: Results on different numbers of generated questions. The horizontal line represents the accuracy achieved with unfixed number of generated questions.

with respect to generating denotative and connotative questions, (2) replace the prompt learning component with random selection. The corresponding results indicate that the instructions regarding question generation objectives are essential to the performance. The prompt learning strategy that implies the correlations between relevant concepts brings further improvement.

Relevance evaluation. Another strategy in our proposed method is to evaluate the relevance of the generated questions by their contribution to the final answer as evidential facts. The results in Table 3 shows that eliminating the relevance evaluation from our proposed method results in a decrease in accuracy. It indicates that the corresponding relevance scores provide additional information for the inquirer to consider the generated questions in the information integration stage.

4.4 In-depth study of question generation

Number of generated questions. Although the number of questions to generate is determined by the inquirer according to the demand of information collection, we conduct complementary experiments to study the impact of the numbers of questions. We vary the total number of generated questions among $[2, 4, 6, 8, 10]$ and present the results in Figure 3. It shows that as the number increases, there is an improvement in accuracy. However, further increments beyond a certain point yields no significant gains. The appropriate number of questions contain the necessary information to answer the target question while mitigating the introduction of redundant information.

Question generation strategies. We conduct experiments to compare with several baseline question generation methods, including slot-filling, free-form generation and LLM generation. Slot-filling generation method defines question templates from different aspects and fills the keywords in the slots to complete the question. We design the templates considering four types of questions following (Ren et al., 2015), including *Object, Number, Color, Location* questions. The question templates are: (1) Is there any [slot] in the image? (2) Is the image about [slot]? (3) What kind of [slot] are there in the image? (4) How many [slot] in the image? (5) What color is the [slot]? (6) Where is the [slot]? We replace the '[slot]' in the template by the major object detected from the image. Free-form generation leverages gated recurrent neural network (GRNN) to predict the output tokens sequentially until hitting the end-of-sentence token, as described in (Mostafazadeh et al., 2016). LLM generation refers to generating questions using LLM without clear guidance about what type of questions to generate. The LLM is asked to generate a few questions given the target question and image caption. For fair comparison, we set the number of gener-

Target question: What style of dress is that man wearing?
Caption: A man wearing a tie and glasses sitting on a bench.
Inspiring words: tie, person, man, style, sight

Denotative questions
Question: What is the man wearing on his face?
Answer: glasses
Question: What is the man wearing around his neck?
Answer: a tie
Question: Where is the man sitting?
Answer: in front of a wall

Connotative questions
Question: What is the overall appearance of the man?
Answer: he wears glasses and a tie
Question: What is the purpose of wearing a tie?
Answer: to make a good impression on others
Question: What is the significance of wearing glasses?
Answer: it is a sign of a person's age
Question: Is the man alone?
Answer: yes

Base VQA model: a suit and tie ✗ **Ours**: formal ✓

Target question: Guess which object is organic?
Caption: A vase with flowers on a table .
Inspiring words: agriculture, flower, element, color, nature

Denotative questions
Question: What objects are present in the image?
Answer: vase, bird, flowers
Question: Are there flowers in the image?
Answer: yes
Question: Is the vase made of organic material?
Answer: no
Question: Is the table made of organic material?
Answer: no

Connotative questions
Question: Are the flowers real or artificial?
Answer: real
Question: Are the flowers grown for decorative purposes or for agriculture?
Answer: for decorative purpose
Question: What is the purpose of the vase in the image?
Answer: to hold flowers

Base VQA model: the bird ✗ **Ours**: flower ✓

Figure 4: Qualitative examples of the generated denotative and connotative questions, and the selected question-specific inspiring words.

ated questions of the compared methods mentioned above to 6. The results are shown in Table 4.

The results indicate that our proposed question generation strategy can produce more informative questions towards answering the target question. The slot-filling method lacks diversity and may not be applicable in some situations. The objective of traditional free-form question generation methods diverges from generating informative questions to help collecting essential information. Counting on LLMs to generate questions achieves better results due to its strong natural language understanding abilities and intrinsic knowledge, while still lacks clear objectives as guidance.

4.5 Qualitative Results

In Figure 4 we present qualitative examples to further illustrate our proposed method and compare the answers of our method with those of the base VQA model. In the left example, the base VQA model can successfully extract the explicit image contents with respect to the appearance and dressing of the man, but fails to give a conclusive answer regarding the overall style of dressing. In contrast, our proposed method asks both denotative questions about explicit patterns and connotative questions about other relevant clues that help infer the answer. The LLM integrates the collected information and leverages intrinsic knowledge to draw a correct conclusion. In the right example, the base VQA model mistakenly regards the bird-shaped ornament as real bird. Our proposed method both asks denotative questions that directly query about the attribute of the main objects in the image, and asks connotative questions that associate the attribute of organic with relevant concepts like artificial, decorative, agriculture.

5 Conclusion

In this paper we leverage the rich knowledge of LLMs for knowledge-intensive VQA tasks. We propose a learn to ask (L2A) architecture that inspires the LLM to generate questions to collect essential visual information for answering the target questions. We introduce the concepts of denotation and connotation and propose an efficient prompting strategy to guide the question generation with clear objective, which involves both explicit and implicitly related information. Our proposed method can effectively generate high-quality questions and efficiently collect required information without expensive training or annotations. Experiment results show that our proposed L2A approach achieves promising performance on knowledge-based VQA datasets without requiring in-context examples.

Acknowledgement

This work was supported in part by National Science Foundation grant IIS-2007613.

Limitations

In this work we propose a prompting strategy that guides the LLM to generate denotative and connotative questions. However, some of the generated questions may be difficult to answer for the base VQA model. A potential improvement to current method may be considering the capabilities of the base VQA model in the objective of question generation. The generated questions should be ensured to not requiring commonsense knowledge or reasoning. Besides, the evaluation of open-ended questions still remains an open problem, since previous exact-match metrics can not discern the equivalent expressions of the correct answers. Current evaluation metric may underestimate the model capabilities and provide inaccurate training objectives.

References

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

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

Lichang Chen, Jiuhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou. 2023a. Instructzero: Efficient instruction optimization for black-box large language models. *arXiv preprint arXiv:2306.03082*.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. 2023b. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Shizhe Diao, Zhichao Huang, Ruijia Xu, Xuechun Li, LIN Yong, Xiao Zhou, and Tong Zhang. 2022. Black-box prompt learning for pre-trained language models. *Transactions on Machine Learning Research*.

Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Zero-shot visual question answering with language model feedback.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10867–10877.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2963–2975.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2018.

Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Qian. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 4389–4400, New York, NY, USA. Association for Computing Machinery.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6116–6124.

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

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.

Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. 2018. ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8611–8619.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2023. Improving automatic vqa evaluation using large language models. *arXiv preprint arXiv:2310.02567*.

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.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Badri Narayana Patro, Sandeep Kumar, Vinod Kumar Kurmi, and Vinay Namboodiri. 2018. Multimodal differential network for visual question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4002–4012, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

V Chandra Sekhar Rao. 2017. A brief study of words used in denotation and connotation. *Journal for Research Scholars and Professionals of English Language Teaching*, 1(1):1–5.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

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.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

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 (CVPR)*, pages 14974–14983.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Göran Sonesson. 1998. Denotation and connotation. *Encyclopedia of semiotics*, pages 187–191.

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

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

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 2022*, pages 951–967, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shagun Uppal, Anish Madan, Sarthak Bhagat, Yi Yu, and Rajiv Ratn Shah. 2021. C3vqg: Category consistent cyclic visual question generation. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, MMAsia '20, New York, NY, USA. Association for Computing Machinery.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023. Filling the image information gap for VQA: Prompting large language models to proactively ask questions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

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.

Xiaoying Xing, Mingfu Liang, and Ying Wu. 2023. Toa: Task-oriented active vqa. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

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.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

Zhiling Zhang and Kenny Zhu. 2021. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021*, WWW '21, page 3501–3511, New York, NY, USA. Association for Computing Machinery.

Maosheng Zhong, Yi Hu, Lei Liu, and Ruzhan Lu. 2008. A practical approach for relevance measure of inter-sentence. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 4, pages 140–144. IEEE.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Full prompt

We provide an example with full prompts to the large language model (LLM) in the information collection and integration process, as shown in Figure 5. It includes the prompt to generate questions (*i.e.*, **ASK**) and the prompt to obtain the final answer (*i.e.*, **ANSWER**).

A.2 Choices of different LLMs

In the main experiments we implemented the inquirer in our method by gpt-3.5-turbo. To investigate the influence of different LLMs, we conduct ablative experiments using other LLMs, such as GPT-3 (Brown et al., 2020) and open-source LLM LLaMA (Touvron et al., 2023). We compare the results on OKVQA dataset in Table 5. The results show that GPT-3 achieves even better performance than gpt-3.5-turbo. However, the cost of calling GPT-3 API is much more expensive than that of gpt-3.5-turbo. On the other hand, the performance of LLaMA is not as stable as that of the other two models. It occasionally produces invalid output when presented with same prompts. The overall accuracy using LLaMA is 40.6, while the accuracy considering only the valid samples is 51.2, which is comparable to the other models. The experiments indicate that the performance of our proposed method is influenced by the choice of LLMs, and current LLMs often possess the capability to fulfill the task. Besides, the performance may be further improved by designing model-specific prompts to reduce the invalid outputs of LLMs.

LLM	Accuracy
gpt-3.5-turbo	54.1
GPT-3	**55.9**
LLaMA	40.6 / 51.2

Table 5: Comparison of results on OKVQA dataset using different LLMs.

A.3 Results on VQA v2 dataset

We mainly evaluate our method on knowledge-based VQA datasets, where the results demonstrate that by actively asking questions, the model can effectively utilize the rich knowledge embedded within LLMs, as well as efficiently collect the visual information from images. We conduct complementary experiments on VQA v2 dataset (Antol et al., 2015) for further investigation. It is a commonly used VQA dataset with 443,757 training samples, 214,354 validation samples and 447,793 test samples. Due to the limited visit frequency of OpenAI API allowed to the public and the costs of calling the API, we randomly select a subset of 10,000 samples from VQA v2 validation set. As shown in Table 6, our proposed L2A method outperforms the compared zero-shot methods. The results further indicate the efficacy of L2A method in terms of collecting visual information.

Method	Zero-shot	Accuracy
PICa	✗	56.1
PromptCap	✗	74.1
Flamingo	✓	56.3
BLIP-2 FlanT5$_{XL}$	✓	63.1
PNP-VQA	✓	63.3
Img2LLM	✓	60.6
L2A (ours)	✓	70.1

Table 6: Experiment results on VQA v2 dataset.

A.4 Qualitative examples

Figure 6 and Figure 7 present more testing results on OKVQA dataset. The questions require both visual information from the images and commonsense knowledge for reasoning. Our proposed method generates both denotative questions that extract explicit information and connotative questions that seek implicitly related information that help with reasoning about the answers. The introduction of the question relevance provides more clues for the LLM to consider the impact of each question-answer pair to the final prediction.

ASK

Given a question, define its denotative questions and connotative questions. Denotative questions can be either the rephrasing or decomposition of the original question, which extract explicit information. Connotative questions seek implicitly associated information that can help in understanding and answering the original question. Please generate denotative and connotative questions of the target question.
Caption: A group of people standing on top of a truck in front of the capitol building
Question: Where is this building located?
Inspiring words: structure, skyscraper, city, architecture, house

---

ANSWER

Given some related question-answer pairs and their relevance to the target question, integrate the information and give the very short answer to the target question.
Caption: A group of people standing on top of a truck in front of the capitol building
Denotative questions
Question: What is the name of the building in the image?
Answer: The united states capitol building. Relevance: 0.44
Question: What is the location of the building in the image?
Answer: United states capitol. Relevance: 0.49
Connotative questions
Question: What is the purpose of the building in the image?
Answer: It is a government building. Relevance: 0.59
Question: What is the history of the building in the image?
Answer: It is the capitol building. Relevance: 0.46
Target question: Where is this building located? **Answer:**

PREDICTION

Washington D.C.

Figure 5: Exemplar with full prompts to the LLM.

Target question: Which part of this animal would be in use of it was playing the game that is played with the items the man is holding?
Caption: A man holding a dog on his back.
Inspiring words: dog, person, position, risk, adaptability

Denotative questions
Question: What is the animal in the image?
Answer: dog. Relevance: 0.61
Question: What is the game being played?
Answer: frisbee. Relevance: 0.63
Question: What are the items the man is holding?
Answer: a frisbee. Relevance: 0.61

Connotative questions
Question: Is the animal in the image a dog?
Answer: yes. Relevance: 0.62
Question: Is the game being played a risk game?
Answer: yes. Relevance: 0.52
Question: Is the man in the image adaptable?
Answer: yes. Relevance: 0.53

Prediction: the dog's mouth

Target question: Do you think this person has won or lost a game?
Caption: A woman holding a frisbee in a park.
Inspiring words: person, emotion, company, partner, hunt

Denotative questions
Question: What is the person doing in the image?
Answer: holding a frisbee. Relevance: 0.54
Question: Is the person happy or sad?
Answer: happy. Relevance: 0.56
Question: What is the person holding?
Answer: a white frisbee. Relevance: 0.51

Connotative questions
Question: What do you think the person is feeling?
Answer: excited about the frisbee. Relevance: 0.59
Question: What do you think the person is thinking?
Answer: she is going to throw the frisbee into the water. Relevance: 0.69
Question: What is the person's body language saying?
Answer: she is holding a frisbee. Relevance: 0.54

Prediction: won

Target question: What types of clouds are forming in the sky?
Caption: A group of cars on a street at dusk.
Inspiring words: weather, area, sky, place, astrology

Denotative questions
Question: What are the clouds in the sky?
Answer: dark clouds. Relevance: 0.31
Question: What is the weather like?
Answer: it is cloudy. Relevance: 0.60
Question: What is the temperature?
Answer: not enough information. Relevance: 0.21

Connotative questions
Question: What is the significance of the clouds in the sky?
Answer: they indicate the time of day. Relevance: 0.45
Question: What do the clouds in the sky represent?
Answer: a storm. Relevance: 0.56
Question: What do the clouds in the sky mean?
Answer: it's going to rain. Relevance: 0.56

Prediction: cumulus clouds

Target question: What sport is the child playing?
Caption: A young boy swinging a baseball bat.
Inspiring words: person, son, happiness, field, expression

Denotative questions
Question: What is the child doing in the picture?
Answer: swinging a bat. Relevance: 0.51
Question: What is the child holding in the picture?
Answer: a baseball bat. Relevance: 0.59
Question: What is the child wearing in the picture?
Answer: a baseball cap. Relevance: 0.61

Connotative questions
Question: What is the child's expression in the picture?
Answer: he is ready to hit the ball. Relevance: 0.64
Question: What is the child's age?
Answer: not enough information. Relevance: 0.68
Question: Is the child happy?
Answer: yes. Relevance: 0.49

Prediction: baseball

Figure 6: Examples on OKVQA dataset.

Target question: What is the name of this room if it were to be outside house instead of inside?
Caption: A bathroom with a sink and a window.
Inspiring words: dorm, accommodation, use, reside, specify

Denotative questions
Question: What is the name of this room?
Answer: bathroom. Relevance: 0.54
Question: What is the difference between this room and a room outside a house?
Answer: this room is inside a house. Relevance: 0.36

Connotative questions
Question: What is the purpose of this room?
Answer: it is a bathroom. Relevance: 0.61
Question: What do you typically find in this room?
Answer: a toilet, sink, and a mirror. Relevance: 0.63

Prediction: outhouse

Target question: Where might this man work?
Caption: A man in a blue shirt holding a cell phone and a book.
Inspiring words: person, employment, environment, compensation, crew

Denotative questions
Question: What is the man's occupation?
Answer: janitor. Relevance: 0.65
Question: Where does the man work?
Answer: in a call center. Relevance: 0.54

Connotative questions
Question: What time is it in the picture?
Answer: it is not in the picture. Relevance: 0.50
Question: What is the man's daily routine?
Answer: he is at work. Relevance: 0.49
Question: What is the man's social class?
Answer: middle class. Relevance: 0.54

Prediction: in an office

Target question: What did this guy just hit?
Caption: A man playing tennis on a court.
Inspiring words: person, unit, sport, base, hand

Denotative questions
Question: What is the man doing in the picture?
Answer: swinging a tennis racket. Relevance: 0.67
Question: What is the man holding in his hand?
Answer: a tennis racket. Relevance: 0.53
Question: Where is the man?
Answer: on a tennis court. Relevance: 0.59

Connotative questions
Question: What is the man's emotion in the picture?
Answer: he is ready to hit the ball. Relevance: 0.51
Question: What is the man's intention in the picture?
Answer: to hit a tennis ball. Relevance: 0.62

Prediction: a tennis ball

Target question: What is this woman preparing in the photo?
Caption: A woman preparing food in a kitchen.
Inspiring words: woman, wife, picture, emotion, prepare

Denotative questions
Question: What is the woman doing in the photo?
Answer: preparing food. Relevance: 0.66
Question: What is in the photo?
Answer: a woman preparing food in a kitchen. Relevance: 0.67
Question: Who is in the photo?
Answer: a woman. Relevance: 0.51

Connotative questions
Question: What is the woman's expression in the photo?
Answer: happy. Relevance: 0.50
Question: What is the woman's emotion in the photo?
Answer: happy. Relevance: 0.49
Question: What is the woman's body language in the photo?
Answer: she is standing in front of a stove. Relevance: 0.50

Prediction: food

Figure 7: Examples on OKVQA dataset.