

Look before You Leap: Dual Logical Verification for Knowledge-based Visual Question Generation

Xumeng Liu, Wenya Guo*, Ying Zhang, Xubo Liu,
Yu Zhao, Shenglong Yu, Xiaojie Yuan

VCIP, TMCC, TBI Center, College of Computer Science, Nankai University, China
{liuxumeng, guowenya, liuxubo, zhaoyu, yushenglong}@dbis.nankai.edu.cn
{yingzhang, yuanxj}@nankai.edu.cn

Abstract

Knowledge-based Visual Question Generation aims to generate visual questions with outside knowledge other than the image. Existing approaches are answer-aware, which incorporate answers into the question-generation process. However, these methods just focus on leveraging the semantics of inputs to propose questions, ignoring the logical coherence among generated questions (Q), images (V), answers (A), and corresponding acquired outside knowledge (K). It results in generating many non-expected questions with low quality, lacking insight and diversity, and some of them are even without any corresponding answer. To address this issue, we inject logical verification into the processes of knowledge acquisition and question generation, which is defined as LV²-Net. Through checking the logical structure among V , A , K , ground-truth and generated Q twice in the whole KB-VQG procedure, LV²-Net can propose diverse and insightful knowledge-based visual questions. And experimental results on two commonly used datasets demonstrate the superiority of LV²-Net. The code is released at <https://github.com/michelle191/LV2-Net>.

Keywords: Knowledge-based Visual Question Generation, Logical Verification

1. Introduction

Visual Question Generation (VQG) aims to generate meaningful questions about the given images. It has gained significant research efforts in recent years due to its wide applications, such as data augmentation for VQA (Shen et al., 2021), facilitating visual conversation systems (Patil and Patwardhan, 2020), etc. And recently, the knowledge-based visual question generation (KB-VQG), which requires generating questions that need to be answered with outside information other than the image, has attracted more attention (Uehara and Harada, 2023; Xie et al., 2022).

KB-VQG typically includes two key steps: acquiring knowledge from large knowledge bases and generating questions. In view of the highly similar definitions of VQG and KB-VQG, the corresponding approaches are somewhat related. Most notably, since some researchers of VQG find that directly using answers as constraints can largely improve the quality of generated questions (Wu et al., 2022b; Kai et al., 2021), the answers can also be used to aggregate the retrieved knowledge and generate questions in KB-VQG (Xie et al., 2022).

However, these methods mainly focus on utilizing the semantic information from the images and the target answer to generate questions, but few of them, no matter whether outside knowledge is considered, have checked whether the generated questions can correspond to the given constraints. As shown in Figure 1-(i), although the acquired knowledge and the image contain enough infor-

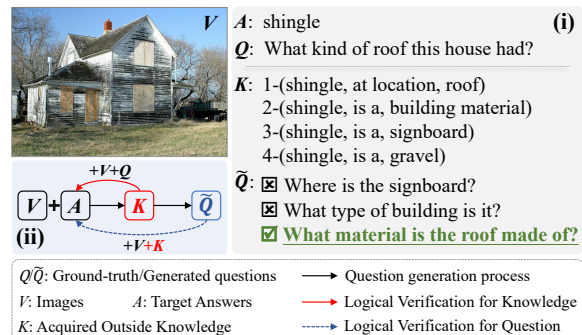


Figure 1: Illustration of answer-aware KB-VQG, where the knowledge entries, represented as (head, relation, tail), are retrieved from a knowledge graph with the visual and the answer.

mation about “shingle”, “roof”, and “house”, the first two generated questions are still far from the ground-truth one. As the lack of logical verification between the elements, the first two questions cannot correspond to the given answer as well.

In this paper, we propose a Dual Logical Verification Network (LV²-Net) to effectively inject logical verification into the two essential steps of knowledge-based visual question generation, including the process of refining the retrieved knowledge and the final generation process. Therefore, LV²-Net consists of two corresponding modules: the Logic-Verified Knowledge Refinement module (LV-KR) and the Logic-Verified Question Generation module (LV-QG). Specifically, LV-KR leverages both the visual and the answer information to refine the external knowledge, working with the logical

* Corresponding author.

verification for knowledge, and producing a logic-enhanced prompt. Then LV-QG utilizes the logic-enhanced prompt to generate visual questions. In addition to keeping consistency with ground-truth questions, LV-QG introduces another logical verification to ensure that the generated questions are logically correct for the whole image-answer question pair. The illustration of the dual logical verification is shown in Figure 1-(ii). The logical verification for knowledge (the red solid line) is used to ensure the logical rationality between the acquired knowledge K and (V, A, Q) pair. And the blue dashed arrow refers to the logical verification for the question, which is to ensure the generated \tilde{Q} can correspond to the target answers well (i.e., the last green sentence in Figure 1-(i)) by checking the logical coherence among \tilde{Q} and (V, A, K) pair.

With the above logical-equipped modules, our approach can improve the quality of generated questions, and promote the logical coherence between the generated questions and other elements. Eventually, experimental results on two benchmark datasets demonstrate that our proposed approach significantly outperforms state-of-the-art methods in KB-VQG and VQG. And the generated questions from our LV²-Net can directly serve as the data augmentation for the task of knowledge-based visual question answering (KB-VQA). The contributions can be summarized as follows:

- We are the first to introduce logical verification in KB-VQG. The proposed LV²-Net can ensure logical correctness in knowledge refinement and question generation.
- The achieved state-of-the-art performance on widely used datasets indicates that questions generated by LV²-Net are of logical correctness to the target answer and high quality.

2. Related Work

2.1. Visual Question Generation

Visual Question Generation (Mostafazadeh et al., 2016a; Fan et al., 2018; Patro et al., 2018) has received more and more research attention in recent years. Early neural VQG methods generate questions solely from the image, namely unconditional VQG (Jain et al., 2017; Mostafazadeh et al., 2016b; Zhang et al., 2016). Such methods can generate a large number of questions, but many of them are not so valuable to be asked or hard to find any correct answers (Bi et al., 2022). Some works implement conditional VQG which incorporates some other auxiliary information from the answer. Yet the answer-type guided results (Krishna et al., 2019a) are still noisy and not so informative. So recently, some works (Wu et al., 2022b; Kai et al., 2021) incorporate the answer as direct supervision for high-

quality question generation. Also, a visual-question answer pair generation method (Yang et al., 2021) measured the consistency of the generated question and answer pairs, but the consistency has not been used to guide better generation. In this paper, we focus on restricting the proposed questions corresponding to the utilized target answers well.

Knowledge-based Visual Question Generation.

Traditional VQG aims to propose visual questions that need to be answered with basic reasoning skills like color recognition. And recently, two specific tasks that require the integration of external information have been defined: Knowledge-Aware VQG (KA-VQG) (Uehara and Harada, 2023) and Knowledge-based VQG (KB-VQG) (Xie et al., 2022). Specifically, KA-VQG refers to asking questions from the target image and a given piece of knowledge, which requires extensive human annotation efforts for the knowledge target. KB-VQG, on the other hand, requires the model to retrieve relevant knowledge and generate appropriate questions independently. In this paper, we focus on KB-VQG due to its greater flexibility when generating better questions with less human labeling.

2.2. Knowledge-based Visual Question Answering

Knowledge-based Visual Question Answering aims to answer questions with external knowledge except for the image content and can be frequently seen in real scenarios. Recent works (Wang et al., 2015, 2017; Narasimhan and Schwing, 2018; Narasimhan et al., 2018; Zhu et al., 2020; Wu et al., 2022a; Marino et al., 2021; Yang et al., 2022; Gui et al., 2021; Lin et al., 2022) incorporated explicit knowledge from various knowledge resources, like ConceptNet (Speer et al., 2017) and Wikipedia (Vrandečić and Krötzsch, 2014). And some other works incorporated implicit knowledge with the ability of pre-trained language models (Yang et al., 2022; Gui et al., 2021; Lin et al., 2022).

As for the datasets, the earliest ones are FVQA (Wang et al., 2017) and KB-VQA (Wang et al., 2015). While the questions and knowledge are relatively trivial or fixed in the above datasets, OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) consist of more flexible and insightful questions and require retrieving knowledge from the external knowledge bases explicitly. In this paper, we follow the previous VQG works (Xie et al., 2022) and implement our method on OK-VQA and A-OKVQA for the KB-VQG model tests.

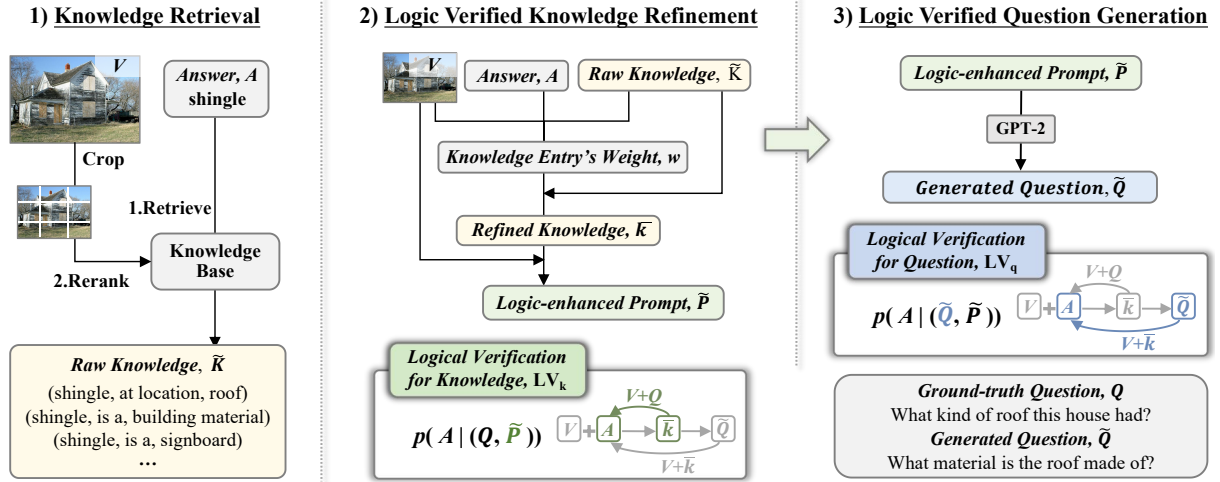


Figure 2: Overview of our proposed Dual Logical Verification Network, in which “LV-KR” and “LV-QG” indicate the logic-verified knowledge refinement module and the logic-verified question generation module, respectively. The dual logic verification LV_k and LV_q are implemented to improve the refinement of outside information and the final generated questions.

3. Method

3.1. Problem Formulation

In this paper, we focus on the KB-VQG task. Formally, given a training dataset $\mathcal{D} = \{V_i, A_i, Q_i\}_{i=1}^{N_{sample}}$, where V_i, A_i, Q_i represent the image, the answer, and the question in the i -th sample. N_{sample} denotes the number of samples. Specifically, V and A are the inputs, Q represents the ground-truth questions, and \tilde{Q} denotes the generated ones, which are the outputs of the task. We use \tilde{K} to represent the retrieved raw knowledge, while \bar{k} signifies the knowledge refined from \tilde{K} . Additionally, a logic-enhanced prompt \tilde{P} , combining both the visual and knowledge information, plays a crucial role in facilitating the generation of compatible and contextually visual questions. Our object is to propose meaningful questions that necessitate both the image and external knowledge.

3.2. Overview

The overview of the model is depicted in Figure 2. The proposed model consists of three components: Knowledge Retrieval Module, Logic-Verified Knowledge Refinement Module (LV-KR), and Logic-Verified Question Generation Module (LV-QG), and the last two modules are trained end-to-end. First, the raw knowledge \tilde{K} is retrieved from ConceptNet (Speer et al., 2017) with both the visual and answer information. Then LV-KR refines the \tilde{K} into \bar{k} with predicted weights to incorporate more related and important information for the final question generation. Also, a logic-enhanced prompt \tilde{P} , combining both the knowledge \bar{k} and the visual in-

formation from V , is aggregated by LV-KR to guide the next question generation. After that, LV-QG generates the final question \tilde{Q} with the prompt \tilde{P} . Meanwhile, the logical verification for knowledge LV_k works on the $\langle A, Q, \tilde{P} \rangle$ to improve the quality of the refined knowledge. And, the logical verification for question LV_q among $\langle A, \tilde{Q}, P \rangle$ works on \tilde{Q} to make it correspond to the target answer. By incorporating these modules, both the knowledge refinement and the question generation procedure are enhanced for better knowledge-based visual question generation.

3.3. Knowledge Retrieval

In this section, we first encode the inputs and then retrieve the raw knowledge. Initially, we implement a pre-trained CLIP (Radford et al., 2021) to encode the answer into a and the image into v_p . Specifically, we encode the image v_p after cropping it into several patches to focus on the detailed information of every visual part. And both the visual and the answer features are used in the retrieval procedure for more expected-question-related knowledge. First, we retrieve the related knowledge entries from ConceptNet, by calculating the inner product between a and the CLIP text embedding of each entity from ConceptNet, and returning the corresponding edges with higher results. This approach provides more flexibility than keyword matching, and the process is facilitated by FAISS (Johnson et al., 2019). After searching, the returned edges are then re-ranked based on the maximum inner product between the embeddings of each image patch v_p and the node embedding on the other side of the edge. Finally, we output

the top N_k pieces as the raw knowledge \tilde{K} .

3.4. Logic-Verified Knowledge Refinement Module (LV-KR)

In this section, LV-KR refines the raw knowledge \tilde{K} into \bar{k} to extract the more accurate external information, then combines the visual and external information into a logic-enhanced prompt \tilde{P} to guide the question generation in the next module. And \tilde{P} is optimized by the logical verification for knowledge LV_k , which can help improve the quality of \bar{k} . By incorporating LV-KR, the model is equipped with the necessary knowledge and logic foundation to propose insightful questions.

3.4.1. Knowledge Refinement

Knowledge encoding. First, we encode the raw knowledge entries into GPT-2 style (Radford et al., 2019), and map it into the scope of the CLIP embedding with a transformer-based reversed mapper innovated by ClipCap (Mokady et al., 2021). Each piece of knowledge in \tilde{K} is represented by the CLIP-style embedding \tilde{k}_{cj} , where j refers to the j -th entry of the raw knowledge.

Knowledge entry’s weight calculating. Then, we predict the weight of each knowledge entry. We encode the whole picture into v_g to leverage the global visual feature. And a query q is calculated with visual and answer information via $q = f^a(v_g \oplus a)$, where f^a is a L2-normalized transformer, and \oplus denotes the concatenate operation. Then, the context weight w^{con} could be obtained with a filter f^{con} :

$$w^{con} = W_{con} f^{con}(q, \tilde{k}_{cj}), \quad (1)$$

where W_{con} are trainable parameters, and the filter f^{con} can calculate the cross-attention between the query and the knowledge piece embedding. After that, the final weight w could be aggregated from w^{con} and the weight for each piece w^{KB} given by ConceptNet:

$$w = softmax(W_w(w^{con} \oplus w^{KB}) + b), \quad (2)$$

where both W_w and b are trainable parameters.

Knowledge refining. With w and \tilde{k}_{cj} , the refined information \bar{k} of the retrieved raw knowledge can be calculated as:

$$\bar{k} = Norm\left(\sum_j^{N_k} w \cdot \tilde{k}_{cj}\right). \quad (3)$$

The information included by \tilde{K} is refined, leaving the useful external knowledge features being the foundation of the next knowledge-based question generation.

3.4.2. Logic-enhanced Prompt Acquisition

In this paper, we designed a logic-enhanced prompt \tilde{P} , which can measure the logical relationship between Q and A , to prompt the final question generation. And LV-KR obtains \tilde{P} by combining the refined knowledge \bar{k} with the whole picture information v_g :

$$\tilde{P} = Mapper(W_P(\bar{k} \oplus v_g)), \quad (4)$$

where W_P are trainable parameters, $Mapper(\cdot)$ (Mokady et al., 2021) can map the CLIP-scope embedding into GPT-2-scope. The logic-enhanced prompt \tilde{P} can provide contextual and logical information to guide better question generation.

3.4.3. Logical Verification for Knowledge

\tilde{P} plays a crucial role in guiding the generation of \tilde{Q} . And to improve the quality of \tilde{P} , we introduce the logical verification for knowledge function (LV_k) on \tilde{P} , Q , and A , and make the plausibility $p(A | (Q, \tilde{P}))$ be close to 1. Meanwhile, as \tilde{P} is the combination of v_g and \bar{k} , LV_k can build up the strong logical relationship between (Q, A, V, \bar{k}) . Specifically, we utilize a sentence template, "Question: Q , reason: \tilde{P} . The answer is ". We fill it and provide it as a prompt for a fixed GPT-2. A tentative answer \tilde{A}_k can be generated from the expected question Q and the logic-enhanced prompt \tilde{P} . $p(A | (Q, \tilde{P}))$ is maximized via the Cross-Entropy Loss between \tilde{A}_k and A . Thus with LV_k , the quality of \tilde{P} , where \bar{k} in fact, can be improved to be more accurate and rational. Enables the model to effectively guide the next generation of \tilde{Q} better.

3.5. Logic-Verified Question Generation Module (LV-QG)

In this part, the final sound question \tilde{Q} would be proposed, guided by the logic-enhanced prompt \tilde{P} , and another logical verification LV_q would be implemented to improve the quality of \tilde{Q} .

3.5.1. Question Generation

In the question generation phase, we generate the final question \tilde{Q} by inputting \tilde{P} as the prompt of a trainable GPT-2. We leverage its language capabilities to generate contextually relevant and logically correct questions with the help of \tilde{P} .

3.5.2. Logical Verification for Question

We utilize a similar template introduced in LV_k , "Question: \tilde{Q} , reason: \tilde{P} . The answer is ", to prompt the generation of \tilde{A}_q . In this round of verification, as our object is maximizing the plausibility

$p(A | (\tilde{Q}, \tilde{P}))$, the generated \tilde{Q} , instead of Q , is sent as the input to a fixed GPT-2. This allows us to evaluate the logical correctness among the generated \tilde{Q} , the verified \tilde{P} (which includes the image information and the knowledge information), and the given answer A . The verification is performed by optimizing the cross entropy value between \tilde{A}_q and A . This process allows us to assess the quality of the generated questions in terms of logical consistency and compatibility with the provided answer, image, and the logic-verified \bar{k} .

3.6. Loss Function

The loss function L incorporates three parts, the question distance loss L_0 , LV_k loss L_k , and LV_q loss L_q :

$$L = \alpha_0 \cdot L_0 + \alpha_k \cdot L_k + \alpha_q \cdot L_q, \quad (5)$$

where α_0 , α_k and α_q are hyper parameters. Specifically, L_0 ensures \tilde{Q} be similar to Q syntactically, L_k and L_q can improve \tilde{Q} semantically. While L_k pays more attention to \bar{k} and L_q pays more attention to \tilde{Q} . And, all the 3 functions are calculated with the Cross-Entropy Loss L_{ce} :

$$L_0 = L_{ce}(\tilde{Q}, Q), \quad (6)$$

$$L_k = L_{ce}(\tilde{A}_k, A), \quad (7)$$

$$L_q = L_{ce}(\tilde{A}_q, A). \quad (8)$$

Conclusively, L_0 improves proposed questions at the word level, L_k enhances the quality of the refined outside knowledge, and L_q aims to build up the strong logical relationship between the generated question and the target answer. By optimizing these components collectively, the model can generate high-quality questions that are both contextually grounded and logically coherent.

4. Experiment

In this section, we first describe the basic information of our experimental setup, then compare our model with the existing methods. Next, the effectiveness of the outside knowledge acquisition part is tested. Moreover, the functionality of LV_k and LV_q is fully experimented and thoroughly analyzed. Last, the quality of LV^2 -proposed questions is proved by a data augmentation test on KB-VQA.

4.1. Experimental Settings

4.1.1. Datasets

Following the existing VQG works (Xie et al., 2022; Kai et al., 2021; Wu et al., 2022b), we test our model on VQA datasets. Specifically, we implement our KB-VQG experiments on two famous

KB-VQA datasets, OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022). **OK-VQA** is the first VQA large-scale dataset that the models have to answer the questions by retrieving the related external knowledge by themselves, rather than incorporating the given and fixed knowledge entries. **A-OKVQA** is a recently proposed KB-VQA dataset that includes questions should be answered with comprehensive reasoning skills and diverse sources of external knowledge.

4.1.2. Implementation Details

We use the pre-trained CLIP (ViT-B/16) to encode the image and the answer. Also, we employ the implementation of (Mokady et al., 2021) for GPT-2, which can ensure both the capability and flexibility when generating sentences, and all the GPT-2-based baselines are compared sharing the same scale fairly. Both the mapper and the reversed mapper are transformer-based and include 8 multi-head self-attention layers and each of them has 8 heads, and the other transformers with 8 layers but 4 heads. We train our model for 30 epochs with a batch size of 20. And we use AdamW (Loshchilov and Hutter, 2019) for optimization with a learning rate of $1e^{-5}$ and 5000 warm-up steps. v_p is calculated after the image being cropped into $N_k = 9$ pieces. For the loss function, we set $\alpha_0 = 0.49$, $\alpha_k = 0.21$, and $\alpha_q = 0.3$. The number of the retrieved knowledge pieces is set to 10 for both the two datasets.

4.1.3. Baselines

In this paper, we compare our model with baselines that utilize different information.

- **ClipCap** (Mokady et al., 2021) can generate image captions with GPT-2 with the image embedding from CLIP. As the similarity between the image caption task and the VQG task, the model can also be trained to generate visual questions that are solely based on the image information.
- **IM-VQG** (Krishna et al., 2019b) encodes the image with ResNet (He et al., 2016) and decodes it with GRU (Chung et al., 2014). The model considers answer categories as auxiliary information.
- **DH-GAN** (Kai et al., 2021) introduce the visual and the answer information, as in "double hints", into the visual question generation process.
- **KVQG** (Xie et al., 2022) is the first knowledge-based visual question generation model.

4.1.4. Metrics

Metrics for automatic assessment. Following previous works, we compare our model with the SoTAs in standard linguistic generation metrics,

		C	B@1	B@2	B@3	B@4	M	R	S
OK-VQA	ClipCap	43.27	25.68	12.79	8.11	5.36	12.41	26.51	10.96
	IM-VQG	30.25	36.47	<u>15.84</u>	8.65	5.04	<u>15.19</u>	36.22	8.03
	DH-GAN	27.27	22.73	9.49	5.92	3.33	7.84	25.61	7.52
	KVQG*	<u>55.38</u>	27.18	15.02	<u>9.87</u>	<u>6.75</u>	13.27	27.17	<u>13.16</u>
	LV ² -Net (ours)	92.17	<u>29.90</u>	18.56	13.15	9.61	15.31	<u>31.94</u>	17.60
A-OKVQA	ClipCap	26.49	26.73	14.43	8.53	4.93	11.50	26.70	8.61
	IM-VQG	22.11	39.30	18.78	10.27	4.85	12.24	38.66	5.64
	KVQG*	<u>40.97</u>	30.56	17.42	<u>10.56</u>	<u>6.27</u>	<u>13.46</u>	30.66	<u>10.72</u>
	LV ² -Net (ours)	60.06	<u>32.11</u>	19.45	12.94	8.64	14.14	<u>33.05</u>	12.84

Table 1: Comparison with existing approaches on two different datasets, OK-VQA and A-OKVQA. The performance is evaluated with various metrics, where “C” denotes CIDEr, “B@1-B@4” denotes BLEU 1 to 4, “M” refers to METEOR, “R” refers to Rouge_L and “S” is SPICE. Also, the “*” in “KVQG*” denotes that we change the LSTM decoder into a GPT-2 decoder to balance the model size difference with our model.

CIDEr (Vedantam et al., 2015), BLEU (1 to 4) (Papineni et al., 2002), ROUGE_L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and SPICE (Anderson et al., 2016). All the automatic metrics are in % from the tables.

Metrics for human evaluation. In this paper, we follow KVQG(Xie et al., 2022) and design 4 metrics: 1) The fluency (F) of the generated questions. 2) Whether the question can be answered (A). 3) Correctness of the retrieved knowledge and their corresponding weight (K). K reflects whether the information retrieved from outside is helpful for proposing expected questions, and whether the calculated weight of each knowledge piece is rational. A higher K indicates better refined knowledge. 4) Logical coherence (L) means whether the generated questions correspond to the target answers. For each metric, we score it with 0, 1, and 2. Higher values on F represent higher fluency, and the same for the other metrics. For each human evaluation test, we randomly select 100 images from OK-VQA. And the performances for each method are evaluated by 10 people with good English skills.

4.2. Q1: Can LV²-Net outperform the existing models?

Main results. As shown in Table 1, our methods can achieve much better results in both of the two datasets on most of the metrics. Even though IM-VQG can achieve better results in B@1 and R, it fails in other metrics. Specifically, for the BLEU metrics, IM-VQG performs better only in B@1. It means that although IM-VQG gives more accurate words, it still fails to generate consecutive phrases and reasonable sentences. Yet the latter abilities on phrases and sentences are more important to propose fluent questions. On the other

hand, Rouge_L can be calculated as:

$$Rouge_L = \frac{(1 + \beta^2) \frac{Lcs}{m} \frac{Lcs}{n}}{\frac{Lcs}{m} + \beta^2 \frac{Lcs}{n}} = \frac{1 + \beta^2}{n + \beta^2 m} \cdot Lcs, \quad (9)$$

where Lcs is the length of the longest common subsequence, m the the word number of the referenced sentence, and n is the length of the generated one, and β is a value set by humans. It can be seen from the equation that with a similar Lcs , the shorter sentence generated by IM-VQG can achieve a higher Rouge_L score. To be short, the simple and shorter sentence generated by IM-VQG can achieve a higher Rouge_L score, which is not what we expect. Thus, compared to baselines, our model can generate more complex questions with deep insight like the ones in the datasets.

Human evaluation. We implement human evaluation on three GPT-2-based models, ClipCap, KVQG*, and LV²-Net. ClipCap works solely from the image information. KVQG* implements the knowledge, and our LV²-Net leverages more refined knowledge and two logical verification.

As shown in Table 2, LV²-Net can achieve good performance on most of the metrics. However, our model is slightly lower than ClipCap on F. That is because the questions generated by ClipCap are relatively simple, even though they are knowledge-based questions. For example, ClipCap is prone to ask about the breed of the animal shown in the picture, but our model tends to ask more complicated and diverse questions like the gestation period, the habitat, and so on. It is obvious that the former questions asked by ClipCap, like 'What is the breed of the animal', are simple and fixed, which are easy to be uttered fluently with high answerability (A). Yet the latter questions from our model are more comprehensive. Like the questions from the datasets, such questions require more knowledge equipped to answer all of them.

On the other hand, the huge progress on K and L shows that our proposed knowledge method and

	F	A	K	L
ClipCap	1.97	<u>1.73</u>	-	0.49
KVQG*	1.92	1.64	<u>0.23</u>	<u>0.66</u>
LV ² -Net	<u>1.94</u>	1.76	1.22	1.33

Table 2: Human evaluation results of the generated questions for the OK-VQA dataset.

	C	B@4	M	R	S
#0	42.37	5.30	12.16	26.17	10.64
#5	<u>90.63</u>	<u>9.52</u>	<u>15.29</u>	<u>31.78</u>	<u>17.52</u>
#10	92.17	9.61	15.31	31.94	17.60
#15	84.29	9.00	14.89	30.73	16.60
#20	88.99	9.43	15.18	31.58	17.04

Table 3: The performance comparison on OK-VQA for different knowledge scales. #x denotes the number of knowledge entries N_k imported for every question generation.

the dual logical verification can work well. Improvement on K means the correct outside information properly retrieved and refined, and the progress on L shows that our generated questions can correspond to the target answers well. And the improvement on A indicates that our model can propose insightful questions that are available to be answered, even though our proposed questions are more complicated. Thus, the essential question generation improvement in real-world scenarios, associated with the huge promotion of utilizing proper outside information, can be seen from the improvement in A, L, and K. And the basic quality of questions is guaranteed from F as well.

Case study. We exemplify cases Figure 3 ①-④ to show the good performance of our model. As ① and ②, it is obvious that the knowledge entries retrieved and refined by our model are more related to the expected question. The cases show the success of LV-KR. As to the ③ and ④, we can notice that ours can generate more target-answer-related questions than the baseline models working with LV-QG. Therefore, our dual logical verification strategy is useful for generating better knowledge-based visual questions indeed.

4.3. Q2: May the knowledge part improve the final performance?

Role of the knowledge scale. We can see from Table 3 that knowledge plays a really important role. The model can perform better when introducing more external information, yet too much information with noise would put huge pressure on the refinement part. Thus the model performs the best when 10 pieces of related knowledge are

	C	B@4	M	R	S
LV ² -Net	92.17	9.61	15.31	31.94	17.60
w/o K	42.37	5.30	12.16	26.17	10.64
K_{token}	43.04	6.13	12.30	27.00	9.03
K_{avg}	84.45	8.90	14.70	31.08	16.56
K_{KVQG}	40.89	5.13	12.08	25.90	10.54

Table 4: Results of different knowledge refinement methods on the OK-VQA dataset.

LV _k	LV _q	C	B@4	M	R	S
✓	✓	92.17	9.61	15.31	31.94	17.60
✗	✓	78.93	8.49	14.69	30.90	15.67
✓	✗	<u>84.69</u>	<u>8.97</u>	<u>14.87</u>	<u>30.80</u>	<u>16.48</u>
✗	✗	70.71	7.44	13.64	29.50	14.89

Table 5: The performance of question generation influenced by the dual logical verification.

✓ denotes our pipeline works with this verification, and ✗ refers to work without it.

retrieved for every proposed question.

Effectiveness of the knowledge retrieval and refinement method. As the ablation results shown in Table 4, it can be noticed that the method plays an important role in better knowledge-based visual question generation.

- **w/o K** is LV²-Net without any external knowledge input. It shows that the knowledge retrieved does help propose knowledge-based questions.

- K_{token} concatenates the words in all retrieved knowledge entries into a paragraph and leverages the token-style embedding from GPT-2 directly without refinement. The performance of K_{token} is limited by the long length of the input paragraph severely.

- K_{avg} aggregates all the knowledge entries with the same weights. Our LV²-Net outperforms K_{avg} in all metrics, which means that the weight is useful for knowledge refinement.

- K_{KVQG} replaces the raw knowledge by the retrieval method from KVQG. K_{KVQG} is retrieved solely with the recognized image objects.

We notice that the K_{KVQG} results are even lower than the **w/o K**. That is because K_{KVQG} cannot cooperate with our refinement process well. And the information in K_{KVQG} is relatively simple and can be easily told by the pre-trained GPT-2. Thus, the introduction of K_{KVQG} with low quality even introduces extra noise, resulting in the low quality.

4.4. Q3: How can the dual logical verification really work?

Improvement concluded by automatic and manual metrics. To evaluate the contribution of each logical verification, we set up another two ablation







	①	②	③	④																							
																											
Q	What kind of store is across the street?	Where would you find these items?	What is this type of vehicle called?	Where would it be located in a house?																							
ClipCap	What is the name of the type of jacket the man wearing?	What type of computer is this?	What kind of truck is this?	What is this device used for?																							
KVQG	K (bicycle, at location, garage)	(tv, is a, abbreviation for television)	(truck, is a, vehicle)	(toilet, at location, bathroom)																							
Q	What kind of business is this?	Is this a new or old computer?	What kind of truck is this?	Name the type of ceramic used to make this toilet in this picture?																							
Ours	K (bodega, is a, shop)	(computer, at location, office)	(truck, is a, vehicle)	(toilet, at location, bathroom)																							
Q	What kind of store is this?	Where would you typically use this type of device?	What kind of vehicle is this?	What part of the house would this be located in?																							
	⑤		⑥																								
		Q: What kind of bear? K: (teddy bear, synonym, teddy)		Q: What vegetable is shown? K: (broccoli, is a, cruciferous vegetable)																							
A: teddy	<table border="1"> <thead> <tr> <th></th> <th>w</th> <th>\tilde{Q}</th> </tr> </thead> <tbody> <tr> <td>w/o LV_q&LV_q</td> <td>0.10</td> <td>What beed of dog is this?</td> </tr> <tr> <td>w/o LV_q</td> <td>0.17</td> <td>What beed of dog is this?</td> </tr> <tr> <td>ours</td> <td>0.22</td> <td>What beed of bear is this?</td> </tr> </tbody> </table>		w	\tilde{Q}	w/o LV _q &LV _q	0.10	What beed of dog is this?	w/o LV _q	0.17	What beed of dog is this?	ours	0.22	What beed of bear is this?	<table border="1"> <thead> <tr> <th></th> <th>w</th> <th>\tilde{Q}</th> </tr> </thead> <tbody> <tr> <td>w/o LV_q&LV_q</td> <td>0.05</td> <td>What food is this?</td> </tr> <tr> <td>w/o LV_q</td> <td>0.10</td> <td>What is the green vegetable on top of this broccoli salad?</td> </tr> <tr> <td>ours</td> <td>0.15</td> <td>What are the veggies on the plate?</td> </tr> </tbody> </table>		w	\tilde{Q}	w/o LV _q &LV _q	0.05	What food is this?	w/o LV _q	0.10	What is the green vegetable on top of this broccoli salad?	ours	0.15	What are the veggies on the plate?	A: broccoli
	w	\tilde{Q}																									
w/o LV _q &LV _q	0.10	What beed of dog is this?																									
w/o LV _q	0.17	What beed of dog is this?																									
ours	0.22	What beed of bear is this?																									
	w	\tilde{Q}																									
w/o LV _q &LV _q	0.05	What food is this?																									
w/o LV _q	0.10	What is the green vegetable on top of this broccoli salad?																									
ours	0.15	What are the veggies on the plate?																									

Figure 3: ① to ④ are cases for comparing with some baselines, ⑤ to ⑥ are cases showing the effectiveness of LV_k and LV_q. w/o LV_k&LV_q denotes our model working without the dual verification, and w/o LV_q means working with LV_k but without LV_q. We show the most representative knowledge piece from the retrieved raw knowledge denoted as K in the figure. The samples are selected from OK-VQA.

LV _k	LV _q	K	L
✓	✓	1.22	1.33
✗	✓	1.04	1.08
✓	✗	1.10	1.03
✗	✗	1.00	0.92

Table 6: An experiment of how the two verification affect the process of question generation. The quality of knowledge and logic are shown.

experiments shown in Table 5 and Table 6. We can see from the automatic metrics (C, B@4, M, R, and S) that both the two logical verification can improve the model performance separately, and moreover, they can work well together to achieve even higher results. Such results show the success of our dual logical verification training strategy. And more sophisticated conclusions could be seen from the human manual K and L in Table 6. That the difference in K between (line3, line4) and also (line1, line2) shows the capability of LV_k. It means that LV_k can help work on more refined external knowledge information. Similarly as the difference in L between (line2, line4) and also (line1, line3), it proves that LV_q can improve the logical coherence between the generated question and the given target answer. And both K and L are improved more when LV_k and LV_q work together.

Detailed cases. Cases ⑤ and ⑥ for the dual logical verification are shown in Figure 3, the weight

of the most ideal retrieved knowledge piece is improved with LV_k. But still, the questions proposed solely with LV_k suffer from the low quality. Hence, LV_q is added to improve the logical coherence, which in return promotes the ideal knowledge entry's weight as well. And the expected and insightful questions are proposed with LV_k and LV_q working together.

4.5. Q4: Could LV²-Net work as the data augmentation for KB-VQA?

To verify the quality of our generated questions, we add them to the training set of OK-VQA for KB-VQA data augmentation. Specifically, we train BUTD (Anderson et al., 2018) and BAN (Kim et al., 2018) with the same experimental setting, and test the answering accuracy. Both the answering processes are aided by the knowledge entries extracted with the image objects followed KVQG. As the data shown in Table 7, the performance of both of the two models can be improved. And the more capable BAN, especially, achieve 5.55% higher after being augmented. Therefore, LV-QG can propose insightful questions corresponding to the target answer and help the downstream applications.

5. Conclusion

In this paper, we propose a model named Dual Logical Verification Network for the task of knowledge-based visual question generation. The model can

	Original	Augmented
BUTD+K _{KVQG}	9.96%	10.33%
BAN+K _{KVQG}	14.30%	19.85%

Table 7: Data augmentation for OK-VQA. The results in the table show the VQA accuracy before and after augmented by LV²-Net.

retrieve more related retrieved knowledge, and work with dual logical verification, which can help better knowledge refinement and more insightful visual question generation. Our model shows superior performance on two datasets compared to the previous baselines with comprehensive and abundant experimental results.

6. Acknowledgement

This research is supported by the National Natural Science Foundation of China (No. 62302243, 62272250, 62077031), the Natural Science Foundation of Tianjin, China (No. 22JCQNJC01580, 22JCQJC00150), and the Fundamental Research Funds for the Central Universities, Nankai University (63231149).

7. Bibliographical References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV*, volume 9909, pages 382–398.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086.
- Satanjeev Banerjee and Alon Lavie. 2005. ME-TEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, pages 65–72. Association for Computational Linguistics.
- Chao Bi, Shuhui Wang, Zhe Xue, Shengbo Chen, and Qingming Huang. 2022. Inferential visual question generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4164–4174.
- Junyoung Chung, Çağlar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.
- Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054.
- Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.
- Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6485–6494.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Shen Kai, Lingfei Wu, Siliang Tang, Yueting Zhuang, Zhuoye Ding, Yun Xiao, Bo Long, et al. 2021. Learning to generate visual questions with noisy supervision. *Advances in Neural Information Processing Systems*, 34:11604–11617.
- Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*, pages 1571–1581.
- Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019a. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2018.
- Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2019b. Information maximizing visual question generation. In *CVPR*, pages 2008–2018. Computer Vision Foundation / IEEE.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Re-vive: Regional visual representation matters in knowledge-based visual question answering. *arXiv preprint arXiv:2206.01201*.

- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.
- Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14111–14121.
- Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016a. Generating natural questions about an image. arXiv preprint arXiv:1603.06059.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016b. Generating natural questions about an image. arXiv preprint arXiv:1603.06059.
- Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. Advances in neural information processing systems, 31.
- Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In Proceedings of the European conference on computer vision (ECCV), pages 451–468.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL, pages 311–318. ACL.
- Charulata Patil and Manasi Patwardhan. 2020. Visual question generation: The state of the art. ACM Computing Surveys (CSUR), 53(3):1–22.
- Badri N Patro, Sandeep Kumar, Vinod K Kurmi, and Vinay P Namboodiri. 2018. Multimodal differential network for visual question generation. arXiv preprint arXiv:1808.03986.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In ICML, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.
- Kai Shen, Lingfei Wu, Siliang Tang, Yuet-ing Zhuang, Zhen He, Zhuoye Ding, Yun Xiao, and Bo Long. 2021. Learning to generate visual questions with noisy supervision. In Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 11604–11617.
- Kohei Uehara and Tatsuya Harada. 2023. K-vqg: Knowledge-aware visual question generation for common-sense acquisition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 4401–4409.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In CVPR, pages 4566–4575. IEEE Computer Society.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78–85.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. IEEE transactions on pattern analysis and machine intelligence, 40(10):2413–2427.
- Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. arXiv preprint arXiv:1511.02570.
- Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022a. Multi-modal answer validation for knowledge-based vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2712–2721.
- Lingfei Wu, Lei Yu, Chen Wang, and Dakuo Wang. 2022b. Visual question generation with answer-awareness and region-reference. US Patent App. 17/163,268.
- Jiayuan Xie, Wenhao Fang, Yi Cai, Qingbao Huang, and Qing Li. 2022. Knowledge-based visual question generation. IEEE Transactions on Circuits and Systems for Video Technology, 32(11):7547–7558.

Sen Yang, Qingyu Zhou, Dawei Feng, Yang Liu, Chao Li, Yunbo Cao, and Dongsheng Li. 2021. Diversity and consistency: Exploring visual question-answer pair generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1053–1066.

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.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2016. Automatic generation of grounded visual questions. arXiv preprint arXiv:1612.06530.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. arXiv preprint arXiv:2006.09073.

8. Language Resource References

Marino, Kenneth and Rastegari, Mohammad and Farhadi, Ali and Mottaghi, Roozbeh. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge.

Schwenk, Dustin and Khandelwal, Apoorv and Clark, Christopher and Marino, Kenneth and Mottaghi, Roozbeh. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. Springer.

Speer, Robyn and Chin, Joshua and Havasi, Catherine. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.