K-VQG: Knowledge-aware Visual Question Generation for Common-sense Acquisition

Kohei Uehara
The University of Tokyo
Tokyo, Japan
uehara@mi.t.u-tokyo.ac.jp

Tatsuya Harada
The University of Tokyo / RIKEN
Tokyo, Japan
harada@mi.t.u-tokyo.ac.jp

Abstract

Visual Question Generation (VQG) is a task to generate questions from images. When humans ask questions about an image, their goal is often to acquire some new knowledge. However, existing studies on VQG have mainly addressed question generation from answers or question categories, overlooking the objectives of knowledge acquisition. To introduce a knowledge acquisition perspective into VQG, we constructed a novel knowledge-aware VQG dataset called K-VQG. This is the first large, humanly annotated dataset in which questions regarding images are tied to structured knowledge. We also developed a new VQG model that can encode and use knowledge as the target for a question. The experiment results show that our model outperforms existing models on the K-VQG dataset. Our dataset is publicly available at https://uehara-mech.github.io/kvqg.

1. Introduction

Asking questions about what one sees is an important skill for humans, especially for children, to learn new visual concepts. By asking questions, children can learn new knowledge about the world more efficiently than by learning passively from a given sources (e.g., textbooks). Therefore, it is essential to develop systems that can ask questions about what they see and acquire new knowledge for machine intelligence in the real-world.

Visual Question Generation (VQG) is a research field that aims to give machines an ability to ask questions about images. VQG was initially studied as a task that simply uses an image as input and generates a question related to the image [19]. Recent research on VQG has focused on the way to provide information about the target of a question to the VQG model. Existing studies have used possible answers [14, 15], answer types [11, 25], answer categories [24], and question types [6] as target information.

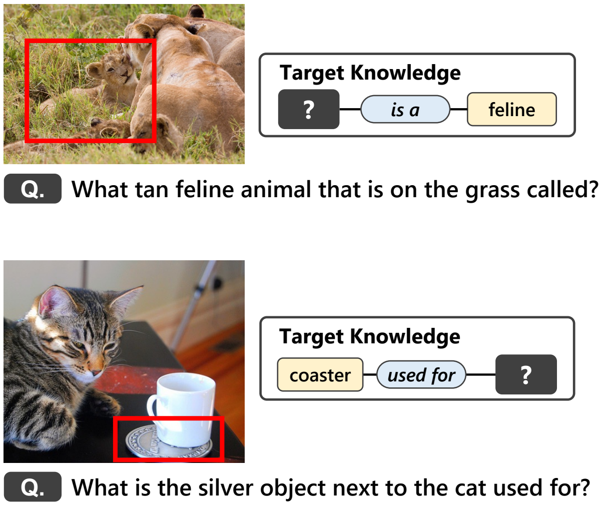

Target Knowledge
?
is a
feline
Q.
What tan feline animal that is on the grass called?
Target Knowledge
coaster
used for
?
Q.
What is the silver object next to the cat used for?

Figure 1. We proposed a new dataset and task in which the model is required to generate a target-knowledge aware question for a given image. In this task, the model is given a knowledge triplet with a missing part and is expected to generate a question that can complement the missing part.

However, when using the answer as a condition, the answer to the question must be known before generating the question. Since questions are usually asked without knowing the answer, such a problem setting is unnatural. Other target information, such as question types, control only the rough target of the question and cannot be used for acquiring specific knowledge.

To establish a more natural and practical setting for VQG, we introduce **K-VQG**, which is a task that utilizes *target knowledge*, i.e., knowledge to be obtained by the question, as target information. Following previous studies on structured knowledge [23, 9], we represent knowledge as a triplet of three words or phrases, i.e., <head, relation, tail>. Specifically, the model takes an image and *masked target knowledge*, which is a knowledge triplet in which a part of the triplet is masked out as input and generates a question such that the answer will be helpful in complementing the missing part. This method allows for fine-grained question control without the need to know the expected answers in advance. For example, in the top example of Figure 1, the target knowledge is <lion, is a, feline>, and the masked target knowledge is <*something*, is a, feline>. The expected output would then be a question related to the knowledge and whose answer would be "lion", e.g., "What tan feline animal that is on the grass called?" On the other hand, the question like "What animal in the image is the top of the food chain?" is indeed a question whose answer will be "lion", but it is not related to the target knowledge.

Since there is no dataset with the necessary annotations (e.g., images, questions, and knowledge triplets), we constructed a new dataset called the **K-VQG dataset**. Our K-VQG dataset is the first VQG dataset that is common-sense aware, human-annotated, and large-scale.

To solve K-VQG task, it is necessary to develop a model that can understand the visual information of an image and masked target knowledge information simultaneously. Existing methods for VQG consider only simple target information, such as answers and categories, and thus cannot handle complex auxiliary information, such as a knowledge triplet. We developed a novel model for K-VQG, which can encode the image and masked target knowledge using a multi-modal transformer based encoder to generate questions. Our contributions are summarized as follows:

- We construct a novel VQG dataset with knowledge annotations called K-VQG.
- We propose a knowledge-aware VQG model that uses a masked knowledge triplet as input.
- We evaluate the performance of the proposed model on the constructed dataset.

2. Related Work

2.1. Knowledge-aware VQA/VQG Dataset

In this section, we introduce Visual Question Answering (VQA) datasets in addition to VQG datasets because many VQG studies use VQA datasets. We summarize the main features of various datasets in Table 1.

The VQA v1/v2 dataset [2, 8] is a widely used dataset in VQG research [15, 14, 10, 11], but it does not contain any knowledge annotations. The FVQA dataset [27], in which the questions are annotated with common-sense triplets, is similar to our own. However, the FVQA dataset is relatively small ($\sim$5K questions), and many of the questions tend to refer primarily to the target knowledge and less to the content of the images. The questions in the FVQA dataset often refer to the image with only phrases like "...in the image". Such questions can be easily generated without understanding the content of the image and are therefore unsuitable for use in VQG. The OK-VQA dataset [17] is intended to be a VQA dataset that requires knowledge and is larger than the FVQA dataset ($\sim$10K questions); however, it lacks annotations on "which knowledge is relevant to the question." The K-VQA dataset [22] is specialized for knowledge of named entities (e.g., "Who is to the left of Barack Obama?"), and its question annotations are template-based, making it less generalizable. The CRIC dataset [7] is a more recently proposed dataset. This dataset is a VQA dataset that contains annotations for common-sense triplet as well as ours. However, this dataset is not annotated by humans, but is a rule-based dataset that automatically generates sentences from scene graph information.

Our dataset has the several advantages over the existing datasets mentioned above: (1) the questions are associated with common-sense knowledge triplet, (2) annotated by humans, (3) bounding box annotations of the question target, and (4) large scale.

2.2. VQG Model

VQG is the task of generating questions associated with images. The earliest VQG model [19] used an RNN model to generate questions using only an image as the input. However, such a model conditioned only on images cannot control the target of a question. Therefore, researchers have been studying ways to control the target of a question by providing additional information. In addition to images, iQAN [14] and iVQA [15] use answers as inputs to generate questions that can produce the desired answers. With these methods, the answer to the question must be known in advance. Since questions are usually asked without knowing the answers, such problem setting is unnatural.

Other methods use categories of answers as conditions for VQG [11, 25]. With these methods, it is not necessary to know the answers to the questions. However, there is a problem that the granularity of the answer categories greatly affect the quality of the control of the question content. Although existing studies [11, 25] use 15 categories, the classification is rather coarse because all answers related to the name of the object are gathered in the "object" category. This means that, when there are multiple objects in an image, it is impossible to control which object should be the target of the generated question.

With our method, the input is a partially masked common-sense triplet. Thus, our method has the advantage of being able to control the target in more detail than the existing VQG models, and it is also easy to apply the acquired information to a knowledge database.

3. K-VQG Task and Dataset

We provide an overview of the K-VQG task, which is a VQG task for knowledge acquisition. In the K-VQG task, the model is given a **masked target knowledge** triplet and

	Num. of Q	knowledge type?	structured knowledge?	target bounding box?	manually annotated?
VQAv2 [8]	1.1M	N/A	✗	✗	✓
FVQA [27]	5,826	common-sense	✓	✗	✓
OK-VQA [17]	14,055	open knowledge	✗	✗	✓
K-VQA [22]	183,007	named entities	✓	✗	✗
CRIC [7]	1.3M	common-sense	✓	✓	✗
K-VQG	**16,098**	**common-sense**	✓	✓	✓

Table 1. Comparison of key features of the major VQG/knowledge-aware VQA datasets. Our dataset is the first manually-annotated VQG dataset that contains knowledge annotations and bounding box annotations.

an **image**, and the model is expected to generate a question that can acquire the **target knowledge**. The masked target triplet is a knowledge triplet in which a part of the question to be answered is masked, e.g., <[MASK], IsA, feline>. By contrast, the target knowledge is a complete triplet in which the masked parts are filled, e.g., <lion, IsA, feline>. For example, the goal of this task is to generate questions from a masked target triplet, such as <[MASK], IsA, feline>, such that "lion" can be obtained as an answer, and knowledge <lion, IsA, feline> can be acquired.

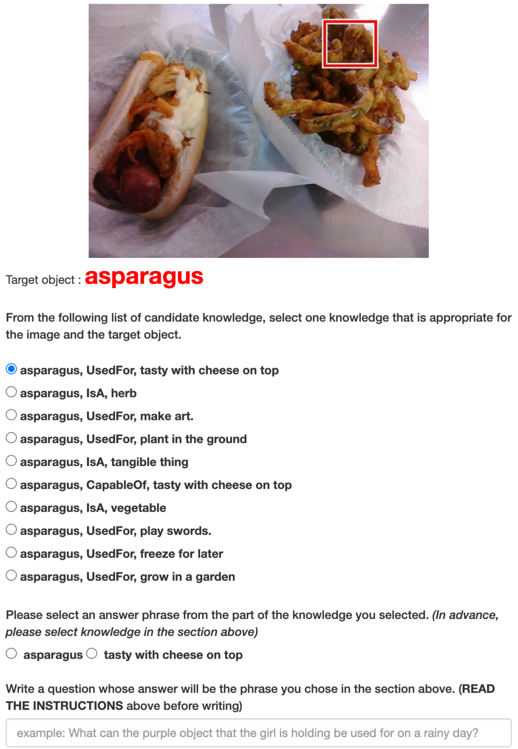
Next, we describe how we construct the K-VQG dataset. Each sample contains the following data: (1) image, (2) question, (3) answer, (4) target knowledge triplet, (5) bounding box of the question target. We asked crowd workers of Amazon Mechanical Turk (AMT)[1] to annotate the data. We sampled the images from the Visual Genome dataset [12], which contains a large and diverse set of object categories, and selected the target object and candidates for the target knowledge (3.1 (a)). We then asked the workers to select one target knowledge and write questions about the image that required the target knowledge to answer (3.1 (b)). We further conduct the question validation process to ensure the quality of the dataset (3.1 (c)).

3.1. Dataset Construction

(a) Knowledge triplet collection. We utilized ConceptNet [23] and ATOMIC$^{20}_{20}$ [9] as the knowledge sources.

ConceptNet is a large-scale knowledge base that contains knowledge collected from several resources. Knowledge in ConceptNet is represented as a triplet of the form <head, relation, tail>, such as <cat, AtLocation, sofa>. ConceptNet contains approximately 34 million triplets and 37 types of relations. Some relations seem to be unnatural targets for questions regarding images, such as *DistinctFrom* or *MotivatedByGoal*. Thus, we selected 15 types of relations that were considered suitable as targets for the questions.

The ATOMIC$^{20}_{20}$ consists of more then 1M knowledge triplets about physical-entity relations (e.g., <bread, ObjectUse, make french toast>), event-centered relations (e.g., <PersonX eats spinach, isAfter, PersonX makes dinner>), and social-interactions (e.g., <PersonX calls a friend, xIntent, to socialize with their friend>). We used only physical-entity relations for our dataset construction because the other relation types were less relevant to the images in the Visual Genome.



Target object : asparagus

From the following list of candidate knowledge, select one knowledge that is appropriate for the image and the target object.

● asparagus, UsedFor, tasty with cheese on top
○ asparagus, IsA, herb
○ asparagus, UsedFor, make art.
○ asparagus, UsedFor, plant in the ground
○ asparagus, IsA, tangible thing
○ asparagus, CapableOf, tasty with cheese on top
○ asparagus, IsA, vegetable
○ asparagus, UsedFor, play swords.
○ asparagus, UsedFor, freeze for later
○ asparagus, UsedFor, grow in a garden

Please select an answer phrase from the part of the knowledge you selected. *(In advance, please select knowledge in the section above)*

○ asparagus ○ tasty with cheese on top

Write a question whose answer will be the phrase you chose in the section above. (**READ THE INSTRUCTIONS** above before writing)

example: What can the purple object that the girl is holding be used for on a rainy day?

Figure 2. Screenshot of the AMT interface (excluding the instruction due to space limitation). The information provided to the worker was displayed at the top of the screen, including the image, target object, and candidate knowledge triplets. Below that, there are sections for the answer phrase selection and writing knowledge-aware questions for the selected answers.

After the above pre-processing, we merged these two knowledge datasets. Then, to remove knowledge that is unrelated to any objects in the images, we queried the en-

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

tity appearing as the head of the knowledge in the Visual Genome object list and removed the knowledge if there was no matching object. Finally, we obtained a total of ∼150K knowledge triplets as candidate knowledge.

(b) Question collection. We show the screenshot of the AMT interface in Figure 2. In order to maintain quality, we selected workers who resided in the U.S. or Canada and had an approval rate greater than 97%. The workers were given the following information: the target image, a bounding box representing the area of the target object (i.e., the head entity of the candidate knowledge), the name of the target object, and a list of candidate knowledge triplets (up to 15). They were asked to write questions with the following steps:

1. From the list of candidate knowledge, select one knowledge that is appropriate for the image and the target object.
2. Select a phrase from the selected knowledge (i.e., head entity or tail entity) to be the answer.
3. Write a question whose answer will be the selected phrase and requires the knowledge the worker have chosen to answer.

We instructed the workers to include a description of the position of the object in relation to other objects in the image, more than simple phrases such as "...in the image". In addition, we instructed to assume that the bounding box is not visible, i.e., phrases such as "surrounded by a red frame" or "with a bounding box" are prohibited.

(c) Question validation. To ensure the quality of the collected questions, we conducted validation of the collected annotations by AMT. We asked workers to evaluate questions with the following criteria: (1) whether the question refers to the visual content of the image, (2) whether the target knowledge is related to the question, (3) whether the target knowledge is related to the image and the target object, (4) whether the question contains typos or grammatical errors, (5) whether the answer is proper for the question. We asked three workers per question for evaluation, and excluded the questions in which all workers unanimously gave negative ratings for any of the evaluation criteria. Note that we evaluated some of the data ourselves in advance, and rejected submissions from workers whose agreement rate with our evaluation was less than 60%, in order to maintain the quality of the evaluation.

3.2. Dataset Statistics

We show the basic statistics of our dataset and two existing datasets in Table 2. From these statistics, we can conclude that our dataset is more challenging for the VQG model. First, our dataset has a longer sentence length for questions and answers than the others. This can also be observed in Figure 3, which shows the distribution of the length of the questions in each dataset. This makes our dataset a more diverse and challenging VQG dataset than the others. In addition, compared with FVQA, our dataset has more non-knowledge words in the questions (3.35 vs. 0.99). This means that most of the questions in FVQA consist of words derived from the knowledge triplet, whereas the questions in our dataset contain many words that are not derived from the knowledge triplet (e.g., words regarding the content of the image). In other words, for the FVQA dataset, the VQG model could generate questions without understanding the content of the images. However, because the questions in the K-VQG dataset require references to the image content, generating proper questions is much more difficult. In Figure 4, we show the word clouds of the most frequent words in the questions and answers in the K-VQG dataset. This indicates that the questions in our dataset pertain to diverse fields (e.g., food, animals, and location). Some examples from the dataset are shown in Figure 5.

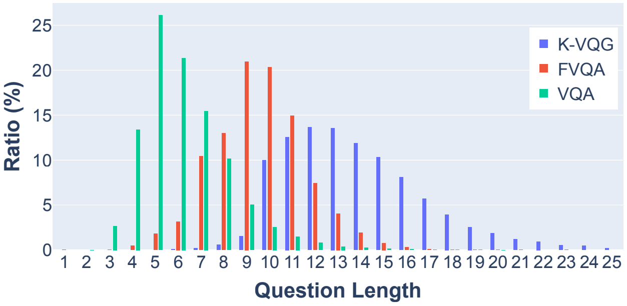
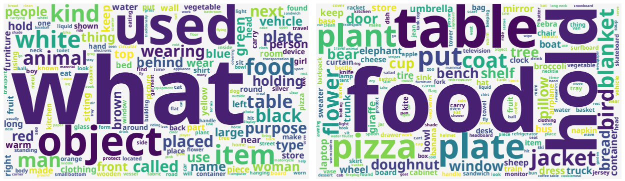


Figure 3. The distribution of question lengths in K-VQG dataset, FVQA dataset and VQA v2 dataset. The K-VQG dataset tends to have longer questions than the others.



(a) Word cloud for questions. (b) Word cloud for answers.

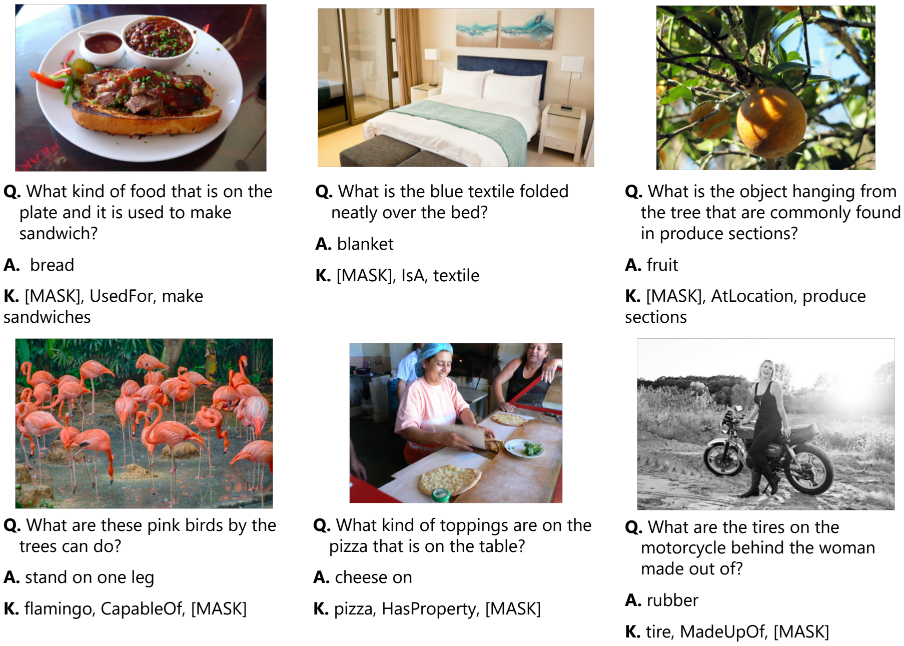
Figure 4. Word cloud for questions and answers. Note that basic stopwords are excluded for the word cloud for questions.

4. Model

The overview of the model is shown in Figure 6. Our model consists of an encoder for the image and target knowledge information and a decoder that uses the output of the encoder to generate the question.

	K-VQG	FVQA [27]	VQAv2 [8]
Num. of questions	16,098	5,826	443,757
– Num. of head answers	11,588	∼4,430	N/A
– Num. of tail answers	4,510	∼1,240	N/A
Num. of images	13,648	2,190	82,783
Num. of unique answers	2,819	1,427	22,531
Num. of unique knowledge	6,084	4,180	N/A
Num. of unique head	527	847	N/A
Num. of unique tail	4,922	2,871	N/A
Average answer length	1.46	1.23	1.10
Average question length	13.88	9.55	6.20
Num. of non-knowledge words in questions	3.35	0.99	N/A

Table 2. **Dataset Statistics.** We compare the K-VQG dataset with FVQA and VQA v2 dataset. *Num. of head/tail answers* indicate the number of answers which is the head or tail entity of the knowledge triplet. Note that the FVQA dataset does not provide such information, and we automatically counted the number. However, because of spelling inconsistencies, we could not obtain an exact count, and thus we used an approximate number here.



Q. What kind of food that is on the plate and it is used to make sandwich?

A. bread

K. [MASK], UsedFor, make sandwiches

Q. What is the blue textile folded neatly over the bed?

A. blanket

K. [MASK], IsA, textile

Q. What is the object hanging from the tree that are commonly found in produce sections?

A. fruit

K. [MASK], AtLocation, produce sections

Q. What are these pink birds by the trees can do?

A. stand on one leg

K. flamingo, CapableOf, [MASK]

Q. What kind of toppings are on the pizza that is on the table?

A. cheese on

K. pizza, HasProperty, [MASK]

Q. What are the tires on the motorcycle behind the woman made out of?

A. rubber

K. tire, MadeUpOf, [MASK]

Figure 5. Example questions and the corresponding images, answers, target knowledge from the K-VQG dataset.

4.1. Encoder

Our encoder uses a pre-trained UNITER [4], which is a multi-modal transformer model that can encode image and text information. In this model, the encoder takes the visual embeddings $\boldsymbol{v}$ of the image and the target knowledge embeddings $\boldsymbol{k}$ as the input and outputs the encoded representation $\boldsymbol{h}$, that is, $\boldsymbol{h} = \mathrm{Enc}(\boldsymbol{v},\ \boldsymbol{k})$.

Visual Embeddings. To obtain visual embeddings $\boldsymbol{v}$, we use a pre-trained Faster R-CNN model [21] and extract region features [1] of the image. Following [4], to provide the positional information of each image region, a seven-dimensional vector representing the coordinates and area of the region was encoded by a linear layer and added to the region image features.

Target Knowledge Embeddings. As described in Section 3, we used partially masked knowledge triplets as input to the model. We treat the masked target knowledge triplet as a sequence of words. Input masked target knowledge is tokenized as asequence of tokens $\boldsymbol{k}\ =$

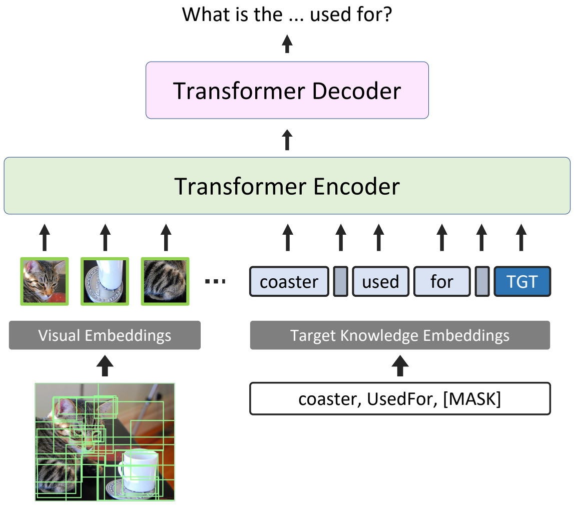


Figure 6. The overview of the model. Our model takes an image and a target knowledge triplet as input, and convert them to fused features by multi-modal Transformer encoder. Then, a Transformer decoder takes the fused features as input and generates a question in an auto-regressive manner.

$\{\boldsymbol{w}_h,\ w_{[\texttt{SEP}]},\ \boldsymbol{w}_r,\ w_{[\texttt{SEP}]},\ \boldsymbol{w}_t\}$. Here, $w_{[\texttt{SEP}]}$ is a special token that indicates the separation of each part, and $\boldsymbol{w}_h,\ \boldsymbol{w}_r,\ \boldsymbol{w}_t$ denote the tokens of the head, relation, and tail phrases, respectively, e.g., $\boldsymbol{w}_h = \{w_{h1},\ w_{h2},\ \ldots,\ w_{hn}\}$. If the head or tail is the masked part, token $\boldsymbol{w}$ is replaced by a special token $w_{[\texttt{TGT}]}$.

4.2. Decoder

The decoder is a module that receives the encoded input image and target knowledge, and outputs the question, that is, $\boldsymbol{q} = \mathrm{Dec}(\boldsymbol{h})$. Following the recent success of transformers in language generation, we developed a transformer-based model for the decoder. Our decoder is also a transformer model, adapted from BART [13], which consists of several transformer blocks, each of which has a multi-head cross-attention and self-attention mechanism. Our model was trained in a teacher-forcing manner by minimizing the negative log-likelihood loss:

$$L_{LM} = -\sum_{n=1}^{|\boldsymbol{q}|} \log P_\theta(\boldsymbol{q}_n \,|\, \boldsymbol{q}_{<n},\, \boldsymbol{h}). \tag{1}$$

5. Experiments

We tested our model and existing methods on the K-VQG dataset. Out of total of 16,098 questions, 12,891 questions were used for training, and 3,207 questions were used for validation. Note that we made sure to split the dataset so that the images used in the UNITER pre-training did not contaminate the validation split of the dataset.

5.1. Implementation Details

Following UNITER, we set the number of Transformer blocks in the encoder and decoder to 12, and the number of hidden units in each block to 768. We initialized the weights of the encoder from the pre-trained UNITER[2]. We used the AdamW optimizer [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. As the learning rate scheduling, we adapted the cosine annealing scheduling, where warm-up steps were set to 10% of the total training steps. The maximum learning rate was set to be 1.0×10^{-5}. We trained the model for 2K steps, which took two hours on 8 × Tesla A100 GPU.

5.2. Baselines

We used several existing methods as the baselines. We did not use any answer-aware VQG models because we did not assume a situation in which the model already knew the expected answer. Thus, we pick VQG models that take images and/or answer categories as input. We automatically annotated the answer categories. If the answer is a word that is the head of the knowledge triplet, we use hypernym dictionary in WordNet [3] to determine the answer category. If the answer is the tail of the triplet, we use this relation as the answer category.

I2Q [19]: The I2Q model is a baseline model based on the approach in [19] that uses only the image as the input and generates a question.

IC2Q: The IC2Q model uses the image and the answer category as the inputs.

V-IC2Q [10, 11]: The V-IC2Q model is a variational autoencoder (VAE) based method, which encodes the answer category and question into a latent space, and decodes the latent vector to generate a question.

IM-VQG [11]: IM-VQG model is another VAE based method. The model is trained to maximize the mutual information between the image, question, and expected answer. Simultaneously, another latent space is learned to encode the answer category, which enables the model to generate questions from only the image and category inputs, without any expected answers.

Input ablation: To demonstrate the importance of input to the model, we performed an input ablation study in which either the image or the target knowledge is excluded from the input to the model (**Ours w/o image**, **Ours w/o knowledge**).

5.3. Evaluation metrics

Following previous VQG research, we used **BLEU** [20], **METEOR** [5], and **CIDEr** [26] as evaluation metrics.

In the K-VQG task, it is also important to evaluate whether the generated questions correctly yield the target

[2] https://github.com/ChenRocks/UNITER

	Question Quality			Knowledge Consistency			
	B-4	M	C	Tri-BLEU	H-Acc	R-Acc	T-Acc
I2Q [19]	11.74	17.05	27.30	4.50	69.69	55.35	1.15
IC2Q	12.37	16.69	31.01	7.97	75.34	58.62	27.91
V-IC2Q [10, 11]	11.78	17.18	28.72	4.70	68.66	55.60	1.53
IM-VQG [11]	11.44	17.07	26.19	4.10	68.07	55.32	1.71
Ours w/o image	17.28	21.06	113.1	61.99	81.95	**83.13**	58.59
Ours w/o knowledge	10.65	16.45	33.92	6.99	65.73	51.01	4.37
Ours	**18.84**	**22.79**	**131.04**	**64.33**	**84.72**	82.44	**66.20**

Table 3. Qualitative results on the K-VQG dataset. The left-side of the table shows the metrics used to evaluate the quality of the questions. Here, B-4, M, and C represent BLEU-4, METEOR, and CIDEr, respectively. The right-side of the table shows the metrics for the knowledge consistency. Tri-BLEU, H-Acc, R-Acc, and T-Acc denote Triplet-BLEU, Head-Acc, Relation-Acc, and Tail-Acc, respectively. For all metrics, higher values are better.

knowledge. To this end, we used the Target Knowledge Parser to predict the masked target knowledge triplet from the generated questions and checked the consistency with the expected knowledge triplet. The Target Knowledge Parser has a similar structure as the K-VQG model. It has a UNITER-based encoder to encode images and questions and a BART-based decoder to recover masked target knowledge. Please refer to the appendix for details of this parser.

We used **Triplet-BLEU** to evaluate the overall agreement between the generated triplets and the ground truth by calculating the BLEU-4 score. In addition, we used **Head-Acc**, **Relation-Acc**, and **Tail-Acc** to evaluate whether each part of the triplet is correct.

5.4. Results and Discussion

We show the experimental results in Table 3. The left side of the table shows metrics of the quality of the generated questions, and the right side shows metrics of whether the generated questions yield the desired knowledge.

Question Quality (vs. baselines) For all metrics used to evaluate the quality of the question, our method outperformed the baselines (Ours vs. others). The baseline method uses only image (I2Q) or image and category (IC2Q, V-IC2Q, IM-VQG) information as input for inference, which suggests that the model has not achieved the ability to sufficiently control the content of the questions to be generated. By contrast, our method directly encodes the target knowledge information and thus succeeds in generating questions with content closer to the ground truth.

Knowledge Consistency (vs. baselines) The right side of Table 3 shows the metrics for knowledge consistency. In terms of Tri-BLEU, which evaluates the overall quality of the generated triplet, our method significantly improves the score compared with other methods. In addition, for part-level accuracy (Head-Accuracy, Relation-Accuracy, and Tail-Accuracy), our method outperformed the other methods. For Head-Accuracy and Relation-Accuracy, our method outperformed the other methods, but the difference was smaller than with the Tail-Accuracy. This is likely due to the fact that the head and relation are often shorter and less diverse than the tail, making it relatively easy to answer correctly even with a conventional method. It should be noted that although tails consist of multiple words, which makes it difficult to generate them correctly, our method can achieve a fairly high accuracy.

Input Ablation. From the input ablation study, it can be seen that when only one of the inputs (image or target knowledge) is used, the performance is worse than when both are used. The performance degradation is particularly noticeable when no target knowledge is input. This may be because target knowledge contains more information about question content control than images. That is, when target knowledge is input, information about what the answer should be is available to the model, whereas when only images are input, such information critical to question content control is not available. These results highlights our claim that the use of desired knowledge as input is important for controlling the content of VQG.

Output Examples. We show several examples of generated questions in Figure 7. In general, our method successfully generates questions that capture the input target knowledge and the content of the images. The bottom three are examples where our model failed to output. From these failed examples, we can see that our model sometimes fails to generate questions when the target object is hidden or too small. In the case of the bottom-right example, the generated question is indeed related to the target knowledge, but the question is about the board itself, not the board material. We believe that further research in methods of encoding image content and knowledge targets will lead to more precise control of question generation.

In addition, we show an example of the output for different inputs in Fig. 8. Comparing w/o image and w/o knowledge, the former outputs more desirable questions because of the more specific information about the target object of the question. This could be the reason why w/o image has

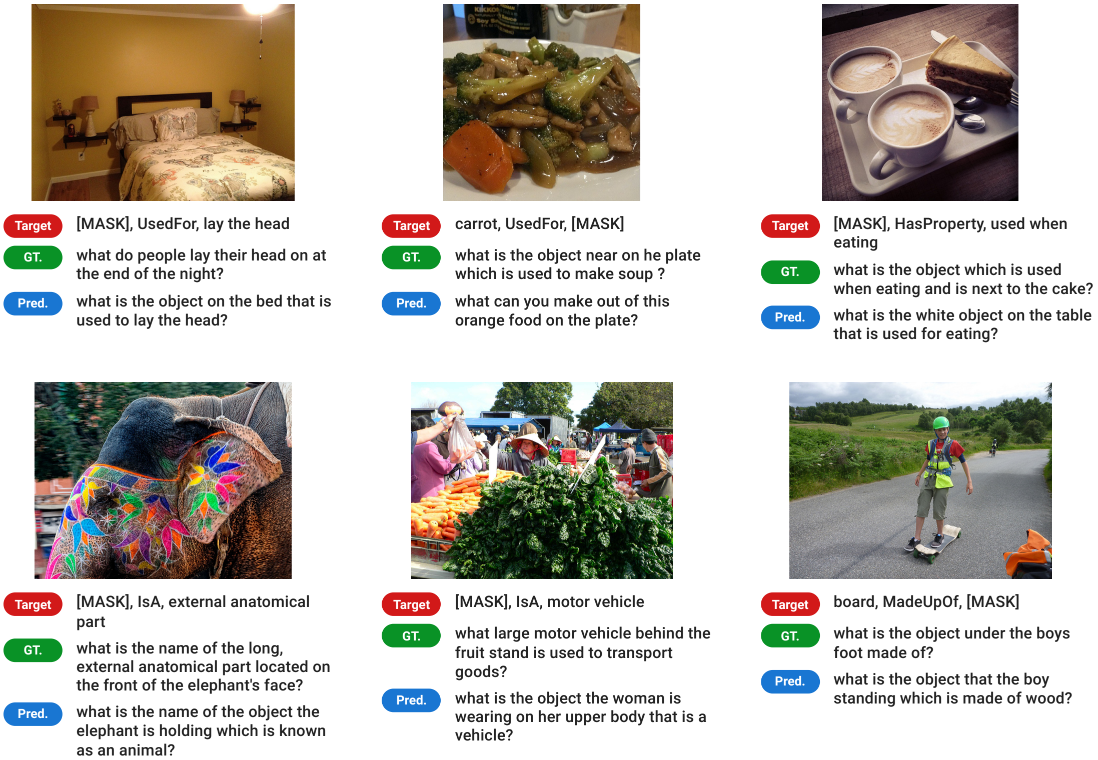

Target
[MASK], UsedFor, lay the head
GT.
what do people lay their head on at the end of the night?
Pred.
what is the object on the bed that is used to lay the head?
Target
carrot, UsedFor, [MASK]
GT.
what is the object near on he plate which is used to make soup ?
Pred.
what can you make out of this orange food on the plate?
Target
[MASK], HasProperty, used when eating
GT.
what is the object which is used when eating and is next to the cake?
Pred.
what is the white object on the table that is used for eating?
Target
[MASK], IsA, external anatomical part
GT.
what is the name of the long, external anatomical part located on the front of the elephant's face?
Pred.
what is the name of the object the elephant is holding which is known as an animal?
Target
[MASK], IsA, motor vehicle
GT.
what large motor vehicle behind the fruit stand is used to transport goods?
Pred.
what is the object the woman is wearing on her upper body that is a vehicle?
Target
board, MadeUpOf, [MASK]
GT.
what is the object under the boys foot made of?
Pred.
what is the object that the boy standing which is made of wood?

Figure 7. Output examples of our method on the K-VQG dataset. We show the input images, target knowledge, ground-truth questions, and generated questions. The first row shows the successful cases, the last row shows the failure cases.

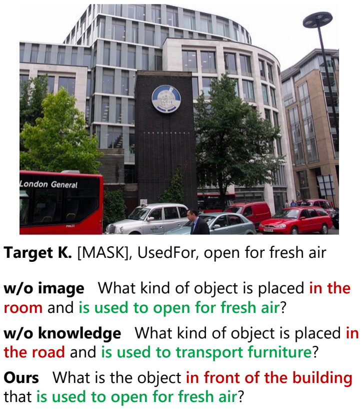

London General
Target K. [MASK], UsedFor, open for fresh air
w/o image What kind of object is placed in the room and is used to open for fresh air?
w/o knowledge What kind of object is placed in the road and is used to transport furniture?
Ours What is the object in front of the building that is used to open for fresh air?

Figure 8. Comparison of model outputs for different inputs. For clarity, phrases related to visual information are colored red, and those related to knowledge information are colored green.

better results overall than w/o knowledge in Table. 3.

6. Conclusion

In this study, we introduce a novel VQG task that uses knowledge as the target of the question. To this end, we constructed a novel knowledge-aware VQG dataset called the K-VQG dataset. The K-VQG dataset is the first large-scale and manually annotated knowledge-aware VQG dataset. We also developed a benchmark model for the K-VQG task. Our experiments demonstrated the effectiveness of our method, while showing some room for improvement.

For future research, our proposed task and dataset have a variety of potential applications. Given the nature of the task, in which the model acquires new knowledge by asking questions, we believe that this task can contribute to the development of learning frameworks, such as human-in-the-loop and learning-by-asking [18]. We expect that this research will lead to the development of a proactive learning system that acquires information about the external world as images and actively learns new knowledge from humans by asking them questions about the images.

Acknowledgements This work was partially supported by JST AIP Acceleration Research JPMJCR20U3, Moonshot R&D Grant Number JPMJPS2011, JSPS KAKENHI Grant Number JP19H01115, and JP20H05556 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo. We would like to thank Naoyuki Gunji, Qier Meng, Hiroaki Yamane, and Yusuke Mori for the helpful discussions.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[3] Francis Bond and Ryan Foster. Linking and extending an open multilingual Wordnet. In *ACL*, Aug. 2013.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020.

[5] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, June 2014.

[6] Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. A question type driven framework to diversify visual question generation. In *IJCAI*, 2018.

[7] Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *arXiv preprint arXiv:1908.02962*, 2019.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017.

[9] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

[10] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, July 2017.

[11] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *CVPR*, June 2019.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, July 2020.

[14] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. Visual question generation as dual task of visual question answering. In *CVPR*, June 2018.

[15] Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, and Changyin Sun. Ivqa: Inverse visual question answering. In *CVPR*, June 2018.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

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

[18] Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens van der Maaten. Learning by asking questions. In *CVPR*, June 2018.

[19] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL*, Aug. 2016.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, July 2002.

[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[22] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, Jul. 2019.

[23] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017.

[24] Kohei Uehara, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Visual question generation for class acquisition of unknown objects. In *ECCV*, September 2018.

[25] Shagun Uppal, Anish Madan, Sarthak Bhagat, Yi Yu, and Rajiv Ratn Shah. C3vqg: Category consistent cyclic visual question generation. In *ACM Multimedia Asia*, 2021.

[26] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, June 2015.

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