Supplementary Materials

Datasets and Preprocessing

Datasets details

In this section, we introduce more details about the datasets used for the evaluation of ConVQG.

K-VQG (Uehara and Harada 2023) is a knowledge-aware VQG dataset. It is the first large, human-annotated dataset in which image-grounded questions are tied to structured knowledge. To build the dataset, knowledge triplets were collected from two sources: ConceptNet and ATOMIC$^{20}_{20}$.

ConceptNet contains ∼34M triples and 37 types of relations, which are not all well-suited for image description; therefore, only 15 types of relations were selected as suitable targets for image-grounded questions. ATOMIC$^{20}_{20}$ contains ∼1M knowledge triplets, among which only physical-entity relations were retained for VQG. Both knowledge bases were then post-processed, giving a total of ∼150K knowledge triplets as candidate knowledge for VQG.

The question collection for K-VQG dataset was performed using Amazon Mechanical Turk (MTurk). The workers were given an image, the bounding box of a target object in the image, the name of the target object, and a list of candidate knowledge triplets. The workers were then asked to write knowledge-aware questions for the image by first selecting an appropriate knowledge triplet and an entity of the knowledge triplet that would be the answer to the question. Finally, an independent phase of question validation was performed on MTurk to ensure the quality of the collected questions.

Each sample in the dataset consists of an image, a question, an answer, a knowledge triplet, and a bounding box of the question target. As a result, K-VQG contains 13648 images and 16098 pairs, related to 6084 knowledge triplets.

In our experiments, we use the same dataset splits as in the original paper.

VQA 2.0 (Goyal et al. 2017) is the most commonly used dataset for VQG evaluation (Krishna, Bernstein, and Fei-Fei 2019; Xie et al. 2021). In particular, VQA 2.0 builds on top of the VQA dataset, which contains 204K images from COCO, 614K free-form natural language questions (3 per image), and over 6M free-form concise answers (10 per question).

Despite the significant progress the VQA dataset enabled in the field, it has been shown that language carries strong priors that can result in good superficial performance (Goyal et al. 2017), even when models do not attend to the visual content. The questions and answers in VQA 2.0 have been carefully curated to alleviate these language biases. The idea is that for every *(image, question, answer)* triplet (I, Q, A) in the VQA dataset, one can find an image I' (similar to I) that results in an answer A' (different from A) to the same question Q.

MTurk is used to collect human-annotated data in two steps: (i) finding the complementary images I', and (ii) collecting answers to the complementary (I', Q) image question pairs. Thus, the VQA 2.0 contains more than 1M *(image, question, answer)* triples, being the largest dataset for VQG evaluation to date.

Works in the literature have used the VQA 2.0 dataset with different train, validation, and test splits. For this reason, we consider two versions of this dataset to report our results: VQA 2.0 small (Xu et al. 2020) and VQA 2.0 large (Krishna, Bernstein, and Fei-Fei 2019). Additional information about these two versions can be found in Section Data preprocessing.

VQG-COCO (Mostafazadeh et al. 2016) was collected by selecting 5,000 images that were also annotated by CQA dataset (Ren et al., 2015) and by VQA (Antol et al., 2015), from the MS-COCO dataset (Lin et al. 2014). The main objective of constructing this dataset is to generate more natural and creative questions. The VQG-COCO dataset contains a total of 2500 training images, 1250 validation images, and 1250 testing images. For each image in the dataset, there are five natural questions and five ground truth captions.

FVQA (Wang et al. 2017) was created for fact-based visual question answering; this means that questions in the dataset need the support of some commonsense knowledge to be answered.

To build the dataset, the authors first collected images from the COCO (Lin et al. 2014) validation set and ImageNet (Deng et al. 2009) test set. Three types of visual concepts were extracted from these images: objects, scene and action. Then, supporting facts were selected from knowledge bases, including ConceptNet (Speer, Chin, and Havasi 2017), DBpedia (Auer et al. 2007), and WebChild (Tandon et al. 2014). Knowledge triplets used from DBpedia concern categories and super-categories; ConceptNet relationships encode commonsense knowledge, while knowledge from WebChild encodes comparative relations. During the question collection phase, human annotators were asked to provide visual questions that required a supporting fact to be answered. FVQA contains 2190 images and 5826 *(question, answer)* pairs. However, questions in this dataset have been criticized for being poorly grounded to the image (Goyal et al. 2017). For this reason, we only use FVQA for the transfer setting of ConVQG. Even though the results need to be taken with a pinch of salt.

More details about the datasets' splits used in this work can be found in Table 6.

Data preprocessing

The detailed data preprocessing pipeline, including dataset splitting, filtering and the creation of textual inputs, is introduced in the following paragraphs. Especially, we describe how to process different types of text inputs (such as knowledge triplets, answers, captions and fact sentences) for different datasets.

VQA 2.0 Small (Answer). Following the preprocessing method in Radial-GCN (Xu et al. 2020), we filter out question types that have "less informative" answers (such as "yes/no"). Although the images for training and test are pre-assigned (Karpathy and Fei-Fei 2015), the filtered question types of Radial-GCN are not publicly available. We try our

Dataset		VQA 2.0 small	VQA 2.0 large	K-VQG	VQG-COCO	FVQA
Train	*QA*	221 708	294 296	12 888	12 500	-
	Img	76 238	80 630	10 915	2 500	-
Test	*QA*	12 940	176 868	3 207	6 250	-
	Img	4 593	40 305	2 730	1 250	-
Total	*QA*	234 648	471 164	16 095	6 250	5 826
	Img	80 831	120 935	13 645	1 250	2 190

Table 6: Summary of datasets used for evaluation of ConVQG. *QA* means the number of question-answer pairs and *Img* means the number of images.

best to make our test set quantitatively similar to previous methods (12,940 QA pairs v.s. 12,938 QA pairs). To do that, we select 28 question types out of 65 in the original annotations according to the previous method (Xu et al. 2018).[11] Then we add two more question types, "what number is" and "how many". For text inputs, the answers are fed into a template: *The answer to the question is [answer]*.

VQA 2.0 Large (Answer). As described in (Krishna, Bernstein, and Fei-Fei 2019), answers in VQA 2.0 dataset are annotated with a set of 15 categories and labeled with the top 500 answers. The top 500 answers consist of 82% of the VQA dataset, resulting in 367K training and validation examples. Because the annotations of VQA 2.0 test set are not available, following the preprocessing method in IMVQG (Krishna, Bernstein, and Fei-Fei 2019), we only use the training and validation set of VQA 2.0 dataset. Keeping the top 500 answers, the processed VQA 2.0 training set is split into an 80-20% train-validation split and the processed validation set is used as the test set.

K-VQG (Knowledge triplet and Answer). For the K-VQG dataset, two types of textual constraints are used to generate questions. For knowledge triplets shown as < subject - predicate - object>, we use templates to generate a short sentence based on the masked knowledge triplet. For instance, < container - CapableOf - [MASK]> is mapped to *container is capable of [MASK]*. The detailed formulating method of 15 relationship categories in the paper can be found in Table **??**. As for the answers as text constraints, we use the same template as VQA 2.0 dataset and turn it into the sentence: *The answer to the question is [answer]*.

VQG-COCO (Caption). We use the same split as previous work (Mostafazadeh et al. 2016; Patro et al. 2018), where there are 2 500, 1 250, and 1 250 images for training, validation and testing. In addition, captions in the annotations are used as text constraints to give a 'focus' for question generation. The dataset is different from others since there is no answer associated with questions. In this case, we use captions as textual guidance to provide some textual cues for question generation. The captions are annotated in the dataset, so they don't require any specific processing.

[11] https://github.com/yikang-li/iQAN/blob/master/data

Relationship	Template
UsedFor	is used for
ReceivesAction	receives action
HasA	has a
Causes	causes
HasProperty	has a property
CreatedBy	is created by
DefinedAs	is defined as
AtLocation	is at location of
HasSubEvent	has
MadeUpOf	is made of
HasPrerequisite	has prerequisite to
Desires	desires
NotDesires	not desires
IsA	is a
CapableOf	is capable of

Table 7: The template to form a sentence based on knowledge triplet.

FVQA (Fact sentence). We use the FVQA dataset as a whole for the transfer experiment, so there is no split for the dataset. In addition, FVQA dataset already has facts as textual cues, hence it doesn't require any further processing.

Metrics Details

As briefly introduced in the main paper, we use a variety of language generation metrics to evaluate and compare ConVQG against competitors: BLEU (Papineni et al. 2002), ROUGE_L (Lin 2004), METEOR (Denkowski and Lavie 2014) and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). They assess the conformity between questions generated by a model and ground truth questions. CIDEr, a TF-IDF-based metric, is the closest to human evaluation for image description compared to the other metrics, according to (Vedantam, Lawrence Zitnick, and Parikh 2015). More details about these metrics are given below:

- **BLEU** (BiLingual Evaluation Understudy): it is obtained by matching text snippets with a set of reference texts. Scores are computed considering the presence of a given text segment in the reference snippets. Therefore, BLEU is a precision-based metric. Several variations of BLEU exist, depending on the number of n-grams to match in the reference text (BLEU-1, BLEU-2, . . ., BLEU-n). BLEU-1 considers only 1-grams, while BLEU-n considers k-grams with k varying from 1 to n.
- **ROUGE_L** (Recall-Oriented Understudy for Gisting Evaluation): it gathers several metrics to evaluate the generated text against the reference. Contrary to BLEU, these metrics are recall-based. In particular, we used the ROUGE_L variant in this work, which measures the longest common sub-sequence between the generated sequence and the reference.
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering): it is classically used for machine translation evaluation. METEOR is based on the harmonic mean

of 1-gram precision and recall, where recall weighs more than precision. It uses exact word matching and the ability to stem and match synonyms.

- **CIDEr** (Consensus-based Image Description Evaluation): it was conceived to evaluate the correspondence between the generated text and the reference, especially for image descriptions. After stemming and representing every text snippet as a set of 1 to 4 grams, CIDEr is computed by first calculating the co-occurrences of these n-grams with reference n-grams. Then, the cosine similarity between n-grams of the generated text and the references is computed, giving less weight to frequent n-grams (which are likely to be less informative).

Experimental Setting Details

Here we give more details about the hyper-parameter settings, mainly about the hyper-parameters in the text decoder and training. For the image input, the image size is set to 480. For the BERT model, the number of hidden layers is 12 and the number of attention heads is 12. For beam search decoding during inference, the number of beams is set to 3. For training, the initial learning rate is 2e-5 and weight decay is set to 0.05.

For more details about the experimental environment, we used torch 1.11.0+cu113 and torchvision 0.12.0+cu113. GPU details are shown in the paper.

Parameter	Value
initial learning rate	2e-5
image size	480
weight decay	0.05
number of beams	3
number of attention heads	12
number hidden layers	12

Table 8: The template to form a sentence based on knowledge triplet.

Quantitative Results

Transfer results on FVQA dataset

Besides the standard visual question generation settings, our model can generate questions for open-domain images and texts using the inference mode. To demonstrate the generalization ability of the proposed ConVQG model, we train it on the K-VQG dataset and test its performance on the FVQA dataset. There were some possible overlaps over images in the K-VQG and FVQA datasets (images from the COCO validation dataset), but the text inputs are annotated differently. More specifically, the text input of each image in the FVQA dataset is a fact sentence rather than a knowledge triplet.

Experimental results can be found in Table 9, where the proposed contrastive ConVQG$_{IT}$ model is compared with the non-contrastive baseline model ConVQG$_B$ in a transfer setting. The contrastive method gains slight improvements on all metrics except METEOR (0.08% on BLEU-4, 0.01% on ROUGE_L and 0.04% on CIDEr).

	BLEU-4	METEOR	ROUGE_L	CIDEr
ConVQG$_B$	2.96	**13.78**	23.67	0.37
ConVQG$_{IT}$	**3.04**	13.77	**23.68**	**0.41**

Table 9: Transfer results on FVQA dataset. Both the baseline method **ConVQG$_B$** and the proposed **ConVQG$_{IT}$** are trained on the K-VQG dataset with knowledge triplets as text input. We report the evaluation results on the whole FVQA dataset.

Comparison method details

This section reports additional results of VQG models from the literature. Tables 10, 11 and 12 present the complete list of results in the VQG-COCO, the K-VQG and the VQA 2.0 datasets, respectively. The comparison method details are as follows.

- I2Q (Mostafazadeh et al. 2016) only uses the image to generate the questions.
- K-VQG (Uehara and Harada 2023) jointly encodes the image and the target knowledge (treated as a sequence of words) using a pre-trained UNITER encoder (Chen et al. 2020b), followed by an autoregressive text decoder to generate the question.
- SAT (Xu et al. 2015) ("Show, Attend and Tell") is one of the earliest works incorporating soft and hard attention into image analysis. This model is built to generate captions, with a CNN as image encoder and an LSTM as decoder.
- DL-VQG (Xu et al. 2018) ("Dual Learning for Visual Question Generation") uses reinforcement learning to jointly perform VQA and VQG.
- IVQA (Liu et al. 2018) implements a conditional question generation model to make use of the answer to generate the question.
- iQAN (Li et al. 2018) is similar to DL-VQG. Same as IVQA, it takes the answers as inputs to help generating the questions.
- IM-VQG (Krishna, Bernstein, and Fei-Fei 2019) (" Information Maximizing Visual Question Generation") uses both the answer and its category to condition the question generation, maximizing the mutual information of the image, the question and the answer. When the dataset has no category, the answer itself is considered as one.
- Radial-GCN (Xu et al. 2020) uses a radial Graph Convolutional Network (GCN) to represent the image content and matches the core information for question generation.
- MOAG (Xie et al. 2021) ("Multiple Objects-Aware Visual Question Generation") is the SOTA method on VQA 2.0, proposing to use answers about multiple objects to generate questions.

Method	BLEU-1	METEOR	ROUGE_L	CIDEr
I2Q (Mostafazadeh et al. 2016)	19.2	19.7	-	-
Creative (Jain, Zhang, and Schwing 2017)	35.6	19.9	-	-
MDN (Patro et al. 2018)	36.0	23.4	41.8	0.51
MC-BMN (Patro et al. 2020)	40.7	22.6	**41.9**	0.50
ConVQG$_{IT}$	**50.2**	**26.4**	40.3	**0.56**

Table 10: Results on the VQG-COCO test sets.

Text constraint	*Method*	BLEU-4	METEOR	CIDEr
Answer	IM-VQG (Krishna, Bernstein, and Fei-Fei 2019)	12.37	16.65	0.39
	ConVQG$_{IT}$	**14.30**	**18.67**	**0.78**
Knowledge Triplet	K-VQG (Uehara and Harada 2023)	18.84	**22.79**	1.31
	ConVQG$_{IT}$	**20.01**	22.66	**1.53**

Table 11: Results on K-VQG dataset.

- C3VQG (Uppal et al. 2021) uses VAE to exploit the visual information for question generation without groundtruth answers.
- Creative (Jain, Zhang, and Schwing 2017) combines variational autoencoders with long short-term memory networks to generate creative questions.
- MDN (Patro et al. 2018) (Multimodal Differential Network) is a multimodal network that uses exemplars for obtaining the relevant context to produce natural and engaging questions by triplet losses.
- MC-BMN (Patro et al. 2020) is a deep Bayesian learning model for probabilistic question generation based on multimodal cues.

Qualitative Results

Diversity. Examples from the VQG-COCO dataset are shown in Fig. 7. Since there is not necessarily an answer associated with the question, captions are used as text inputs. On one hand, it is more difficult to use captions to guide the question generation, since captions are usually the description of the whole image. On the other hand, the uncertainty also brings the diversity of question content. Without obvious guidance for questions, the questions can be anything that is related to image content (captions). The results show that in this case, questions generated by ConVQG can be more natural, creative and diverse. We take them as a special case for ConVQG applications.

Different text inputs. We also show examples from the VQA 2.0 dataset as well as more examples from the K-VQG dataset in Fig. 8 and Fig. 11, respectively. For the VQA 2.0 dataset, the model takes answers as text inputs, while for the K-VQG dataset, text constraints can be answers or knowledge triplets. Comparing those two figures we can see, different text inputs lead to different types of questions. Answers are more precise guidance, where the model can 'guess' the question types from the answers sometimes. For example, if the answer is 'green' then the question probably is about the color of an object in the image. On the other hand, knowledge triplets give external commonsense knowledge that is difficult to obtain from the image directly. By providing this, questions are more informative, diverse and challenging.

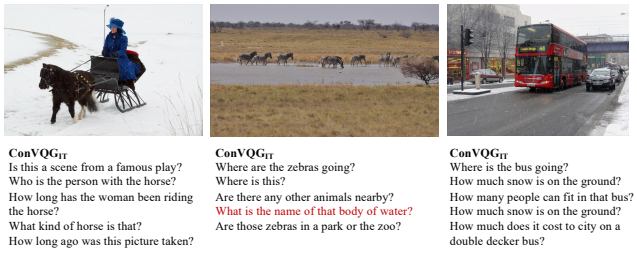


Figure 7: Examples from the VQG-COCO dataset. Since we take captions are constraints in this dataset, which gives more flexibility to the question generation system, the generated questions are more diverse. Red color indicates wrong expressions, not related to the image.

Error analysis. We also provide more examples from the K-VQG dataset, especially some failure cases in Fig. 11. The first two rows show more examples where the generated questions from the proposed ConVQG method can be both image-grounded and text-guided. The last row presents some of the failure cases. For the first and third examples of failure cases (Columns 1 and 3, Row 3), the model generates a question with respect to the text input but adds inappropriate descriptions of image content (e.g. *the ceiling of the room* and *behind the water*). For the first example, the model selects the most likely place where the *fabric* will appear but doesn't pay attention to the image content. For the third example, the model incorrectly detects *water* from the image. For the second failure case (Column 2, Row 3), the model fails to constrain the question by the input text *board is made up of something*, on the contrary, it generates the questions based on the most likely answer *wood*.

Test set	*Method*	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr
Small	SAT (Xu et al. 2015)	49.4	23.1	24.4	53.4	1.65
	DL-VQG (Xu et al. 2018)	50.7	24.4	26.4	55.9	1.88
	IVQA (Liu et al. 2018)	50.2	23.9	**35.7**	55.3	1.84
	IM-VQG (Krishna, Bernstein, and Fei-Fei 2019)	51.3	24.8	26.3	56.3	1.94
	iQAN (Li et al. 2018)	52.6	27.1	26.8	56.9	2.09
	Radial-GCN (Xu et al. 2020)	53.4	27.9	27.1	57.2	2.10
	MOAG (Xie et al. 2021)	58.8	28.1	27.8	60.4	2.39
	ConVQG$_{IT}$	**59.9**	**33.1**	30.0	**62.6**	**2.79**
Large	C3VQG (Uppal et al. 2021)	41.9	10.0	13.6	42.3	0.47
	IM-VQG (Krishna, Bernstein, and Fei-Fei 2019)	**50.1**	16.3	20.6	39.6	0.94
	ConVQG$_{IT}$	45.8	**22.4**	**21.8**	**47.4**	**1.78**

Table 12: Results on VQA 2.0 dataset small/large test set.

Figure 8: Examples from the VQA 2.0 small test set. The answers are used as text inputs.

Human Evaluation

We use MTurk to get human preference in order to evaluate the effect of the contrastive branch of ConVQG.

Selection of examples to evaluate. We asked workers to evaluate 500 examples of the test set of K-VQG dataset, comparing ConVQG$_B$ and ConVQG$_{IT}$ generated questions. From the set of 3207 examples in the test set of K-VQG, we deduplicated images and knowledge triplets. We also removed cases where the baseline model ConVQG$_B$ and the contrastive model ConVQG$_{IT}$ output the exact same questions (155 cases, 4.8% of the test set). Then, we sampled 500 examples, randomly swapping the two questions to avoid bias in the comparison. On top of the two questions to compare and the image, we provide the workers with the knowledge triplet containing the answer to the question; moreover, we highlight in the sentence which section corresponds to the answer, as seen in the examples given to the workers in Fig. 9.

Method	*Votes*
ConVQG$_{IT}$	236
ConVQG$_B$	183
Similar	81

Table 13: Results from MTurk. The vote means the number of times chosen by the annotator in pairwise comparison.

Instructions given to crowd workers. On top of the examples in Fig. 9, we gave detailed instructions to the workers; they can be found in Fig. 10. We list criteria to focus on when selecting the best question relative to the image and the knowledge triplet (which we call *target knowledge* in the instructions). The two main criteria are the grounding of the question to the image and to the knowledge triplet. We specifically asked the workers not to focus on the grammatical correctness of the question to make their choice. Indeed, the difference in architecture and training of the two models we compare should not lead to a significant variation in their ability to generate grammatically correct text; hence, we want the workers to focus on the grounding aspect of the questions. Workers are given the possibility to choose none of the two questions if they consider that the similarity between them is too high to make a meaningful choice. After removing examples where the two questions are identical, however many examples remain where only a few words differ between the two questions. Each worker was given 5 examples per hit. Each hit was only seen by one worker. The workers were pre-selected according to their performance on other tasks.

Overall results. The overall results are shown in Table **??**, where ConVQG$_{IT}$ gets 55 more votes than ConVQG$_B$ among 500 samples.

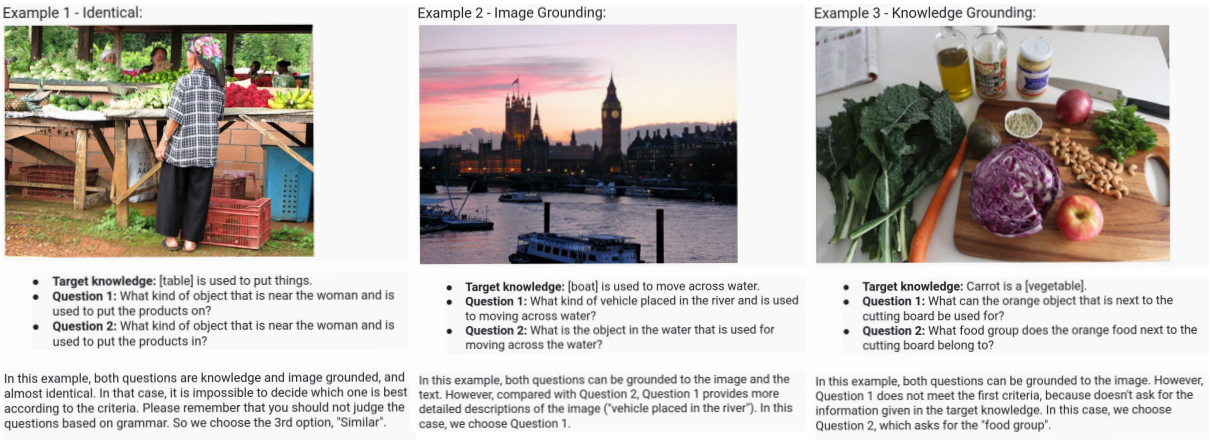


Figure 9: Examples given as instructions for MTurk annotators. We give three different examples: identical, image grounding and knowledge grounding.

Thanks for participating in this HIT!

This task is about evaluating the quality of a question asked about an image.

We provide you with an image, a sentence describing the knowledge we want to ask about (the "target knowledge"), and two questions about the image. The target knowledge is given as a short sentence providing a fact; the element of the sentence between [bracket] is the answer to the question. You are asked to choose the best question according to two criteria:

1. **Grounding to the target knowledge**: is the answer to the question the element between brackets in the target knowledge? Does the question make use of the fact given in the target knowledge?
2. **Grounding to the image**: is the question relevant to the image? Does it require the image to be answered? Does it provide details about what the image is showing?

You are asked to select either **Question 1** or **Question 2** according to these criteria. Sometimes, both questions are very similar; so similar that even given the different criteria, you are unable to decide which one is the best. In that case, you can choose the third option **Similar**. However, we ask you to do your best to pick the best question whenever it is possible.

Additional remarks:

- **If a question is grammatically incorrect:** unless it really prevents you from understanding the question, we ask you not to take this flaw into account when choosing the best question.
- **If none of the two questions ask for the right answer given in the target knowledge:** please choose the one which is the most coherent to the situation in the image.
- **If both questions are poorly grounded or misinterpreted information from the image:** please choose the one which is closest to the fact given in the target knowledge.
- **If both questions are well-grounded to the image and the target knowledge:** please choose the one which provides more detailed information about the image.

Below are some examples to show how to use each criterion:

Figure 10: Instructions given to crowd workers on MTurk.

Input Text: [MASK] is used for surf in the water.

GT: What is the man using to surf in the water?

ConVQG$_{IT}$: What is the object the man is standing on that is used for surfing in the water?

Input Text: [MASK] is a herd animal.

GT: What is the black and white horse which is a herd animal, that is eating the grass called?

ConVQG$_{IT}$: What herd animal is walking through the tall grass and has black and white stripes?

Input Text: [MASK] is used for hold flower.

GT: What is the name of the hollow light green object that is often used to hold flowers?

ConVQG$_{IT}$: What is the green object on the table that is used to hold flowers?

Input Text: [MASK] is used for eat dinner.

GT: What is the object located to the right of the plate and on top of a napkin that is used to eat dinner?

ConVQG$_{IT}$: What is the object on the left of the plate that is used to eat dinner?

Input Text: [MASK] is used for storage.

GT: What kind of object is placed near the bed and is used to storage?

ConVQG$_{IT}$: What piece of furniture next to the bed is used for storage?

Input Text: [MASK] has sub event of hit slopes.

GT: what does the person with the brown jacket have on his left foot that you hit the slopes with?

ConVQG$_{IT}$: What long item does the athlete wear on his foot to hit slopes?

Input Text: [MASK] is a rectangular piece of fabric.

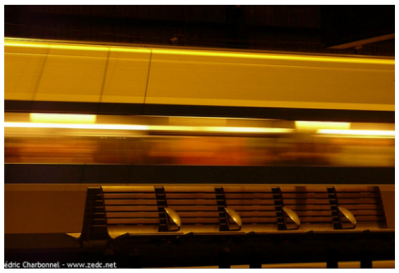
GT: What rectangular piece of fabric hanging outside is used for staying warm?

ConVQG$_{IT}$: What is the name of the object that is a piece of fabric hanging from the ceiling of the room?

Input Text: board is made up of [MASK].

GT: What is the object under the boy's foot made of?

ConVQG$_{IT}$: What is the object that the boy standing on which is made up of wood?



Input Text: {MASK] is used to sit for a bit.

GT: What is the name of the object which people may choose to sit on white waiting for public transportation?

ConVQG$_{IT}$: What is the object behind that water that is used to sit on for a bit?

Figure 11: Additional examples from K-VQG dataset. The first and second rows show examples in which the generated questions are successfully grounded to both image and text. The last row shows some failure cases where the model provides wrong information about image content or text constraints. In the text, green color denotes the sequence that is related to image content, while yellow color denotes the information that is carried by the text input. Red color indicates wrong expressions, not related to the image or the text input.