ConVQG: Contrastive Visual Question Generation with Multimodal Guidance

Li Mi*, Syrielle Montariol*, Javiera Castillo-Navarro*, Xianjie Dai, Antoine Bosselut and Devis Tuia

EPFL, Switzerland
{li.mi, antoine.bosselut, devis.tuia}@epfl.ch

Abstract

Asking questions about visual environments is a crucial way for intelligent agents to understand rich multi-faceted scenes, raising the importance of Visual Question Generation (VQG) systems. Apart from being grounded to the image, existing VQG systems can use textual constraints, such as expected answers or knowledge triplets, to generate focused questions. These constraints allow VQG systems to specify the question content or leverage external commonsense knowledge that can not be obtained from the image content only. However, generating focused questions using textual constraints while enforcing a high relevance to the image content remains a challenge, as VQG systems often ignore one or both forms of grounding. In this work, we propose Contrastive Visual Question Generation (ConVQG), a method using a dual contrastive objective to discriminate questions generated using both modalities from those based on a single one. Experiments on both knowledge-aware and standard VQG benchmarks demonstrate that ConVQG outperforms the state-of-the-art methods and generates image-grounded, text-guided, and knowledge-rich questions. Our human evaluation results also show preference for ConVQG questions compared to non-contrastive baselines.

Introduction

Modern intelligent agents, like chatbots and dialog systems (Ouyang et al. 2022), nowadays achieve (almost) human conversational skills, thanks to the development of large language models (Brown et al. 2020). With the advances in vision-language research, we are now leaning towards visual dialog systems (Das et al. 2017; OpenAI 2023), which should be able to understand and interpret visual scenes and at the same time communicate with users. In this context, they should not only be able to provide answers but also be aware of what they do not know and request complementary information by asking questions about visual content.

Consequently, Visual Question Generation (VQG, (Krishna, Bernstein, and Fei-Fei 2019; Zhang et al. 2017)) becomes a rising area at the intersection of computer vision and natural language processing. VQG agents aim to generate meaningful and engaging questions for visual stimuli such as images. These images often depict multi-faceted scenes, with many salient elements that can be elaborated upon by asking focused questions.

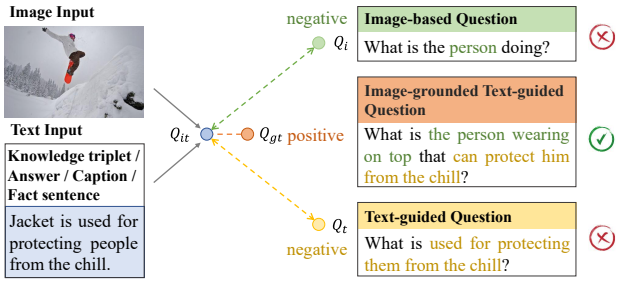


Figure 1: ConVQG at a glance. An image and a text input are processed through a multimodal module , leading to the embedding Q_{it}. Pre-trained modules (detailed in Fig. 2) produce image-only and text-only question embeddings (Q_i and Q_t). A contrastive loss is then optimized to make Q_{it} close to the real question embedding Q_{gt} and far from the single modality ones. By design, ConVQG generates questions that are image-grounded (in green) and that meet the requirements of the text constraint (in yellow).

Early VQG systems tend to generate generic questions not exploiting the rich semantic content of the specific images. For example, the question "*What is the person doing?*" can be asked for any image containing a person. To make the question more focused, existing VQG systems exploit textual constraints, such as expected answers or knowledge triplets, as guidance. However, generating questions that are guided by a textual constraint while enforcing high relevance to the image content remains a challenge, since VQG systems often ignore one or both forms of grounding.

To tackle these challenges, we propose **Con**trastive **V**isual **Q**uestion **G**eneration (**ConVQG**), a system that generates questions that (1) are based on details unique to a specific image, and (2) can be controlled using text to focus on specific objects, actions or concepts. To achieve that, the proposed method uses two modality-specific contrastive objectives to guide the generation of the question. The image contrastive objective drives the question away from a question generated using the image alone. The text contrastive objective drives the question away from one generated using only the textual constraint, enforcing more specific descriptions

*These authors contributed equally.

of the image while providing explicit control over the diversification of the generated questions. The textual constraint format is highly flexible; it can come from the answer to the question, a caption describing the image, or a knowledge triplet associated with an object or an action in the image. The latter, in particular, allows the model to enrich the generated question with image-grounded commonsense knowledge. These elements are found in existing public visual question-answering and question-generation datasets. Together, the two contrastive objectives allow the model to generate a diversified, rich and image-specific set of questions following textual constraints.

Through extensive experiments in standard and knowledge-aware VQG benchmarks, we show that ConVQG consistently outperforms state-of-the-art methods while providing flexibility regarding the type of textual constraints that can be used (answer, knowledge triplet or caption). Additionally, we perform a human evaluation using Amazon Mechanical Turk that shows the effectiveness of the contrastive learning objective to provide image-grounded and text-guided questions.

Related Works

Visual Question Generation. VQG is a particular case of question generation where the goal is to create one or several questions about a given image (Zhang et al. 2017). Early VQG approaches focused on rules or template-based techniques (Vijayakumar et al. 2016; Geman et al. 2015). With the rise of neural networks, VQG was formulated as an image-to-sequence problem, designing an image encoder followed by a decoder to generate questions in natural language (Ren, Kiros, and Zemel 2015; Mostafazadeh et al. 2016; Li et al. 2018; Patro et al. 2018). However, these approaches often lead to poorly image-grounded and generic questions (Xie et al. 2022; Krishna, Bernstein, and Fei-Fei 2019). To avoid generic questions, *text-guided VQG* has emerged, providing systems some guidance to obtain questions with specific properties. The constraint can be either the expected answer (Xu et al. 2020; Xie et al. 2021), a question type (Krishna, Bernstein, and Fei-Fei 2019), specific parts of the image (Vedd et al. 2022) or some external knowledge (Uehara and Harada 2023). In this work, we propose a VQG method to generate questions guided by text inputs (*e.g.*, a knowledge triplet or the expected answer), which, together with our learning objective, ensures that the generated question is image-grounded and knowledge-aware.

Contrastive Learning (CL). The core idea of CL is learning by comparing. Given an anchor, CL defines a positive and a negative distribution, such that samples from the positive distribution (similar inputs) will be pulled together in the latent space while negative samples (dissimilar ones) will be pushed apart. CL has shown impressive performances on self-supervised and supervised learning in computer vision (Chen et al. 2020a; He et al. 2020; Khosla et al. 2020); natural language processing (Oord, Li, and Vinyals 2018; Klein and Nabi 2021), and audio processing (Saeed, Grangier, and Zeghidour 2021) applications. More recently, CL has shown remarkable results for multimodal embedding alignment in vision-language tasks (Radford et al. 2021; Jia et al. 2021). Indeed, contrastive objectives can be exploited to align representations of data pairs from different modalities (*e.g.*, an image and its textual description). In this work, we leverage a contrastive objective to generate questions that consider visual and textual information together by learning a more distinguishable multimodal text-image joint representation from any single modality representation.

Vision-Language Pretraining (VLP). Benefiting from the success of language model pre-training (Devlin et al. 2018; Raffel et al. 2020; Brown et al. 2020) and the recent development of model architectures in the communities (Dosovitskiy et al. 2021), VLP boosts a large amount of vision-language tasks by providing a powerful vision-language joint representation (Gan et al. 2022; Chen et al. 2023). Those representations are usually pre-trained on large-scale datasets (Schuhmann et al. 2021; Lin et al. 2014) using simple objectives such as masked language modelling (Devlin et al. 2018), text-image matching (Radford et al. 2021; Jia et al. 2021) or masked image modelling (Chen et al. 2020b) and can be fine-tuned for various downstream vision-language tasks (*e.g.*, text-image retrieval (Kiros, Salakhutdinov, and Zemel 2014), image captioning (Anderson et al. 2018), visual question answering (Antol et al. 2015)). In this paper, we build our baseline upon one of these models, BLIP (Li et al. 2022), for the powerful abilities provided by VLP. The proposed contrastive objectives serve as one way of tuning models for more readily accessing knowledge, while also distinguishing pure language commonsense from image-grounded ones.

Contrastive Visual Question Generation

This section introduces our proposed visual question generation method, ConVQG, illustrated in Fig. 2. In a nutshell, ConVQG is based on a multimodal encoder-decoder framework, trained in a contrastive way. The multimodal feature is contrasted against negative pairs obtained from single-modality generators to ensure that the generated question can not be obtained from a single modality alone.

Problem Definition

Given an image i, VQG aims at generating a reasonable and pertinent question q. On top of this, the question should meet a given requirement (*e.g.*, reflecting constraints expressed by knowledge triplets or resulting in a given answer), which can be expressed as a text constraint t. The problem is solved by a multi-modal question generation model $p(q|i,t)$, which embeds image and text into a joint embedding and decodes a question based on image content and text constraints.

Architecture

ConVQG is built based on BLIP (Li et al. 2022), which is a large-scale vision-language pre-training pipeline consisting of an image encoder, a text encoder and a text decoder. Nevertheless, our proposed contrastive method can be used with any vision-language model.

Image Encoder. The image encoder is a vision transformer (ViT) (Dosovitskiy et al. 2021). It receives an image i as

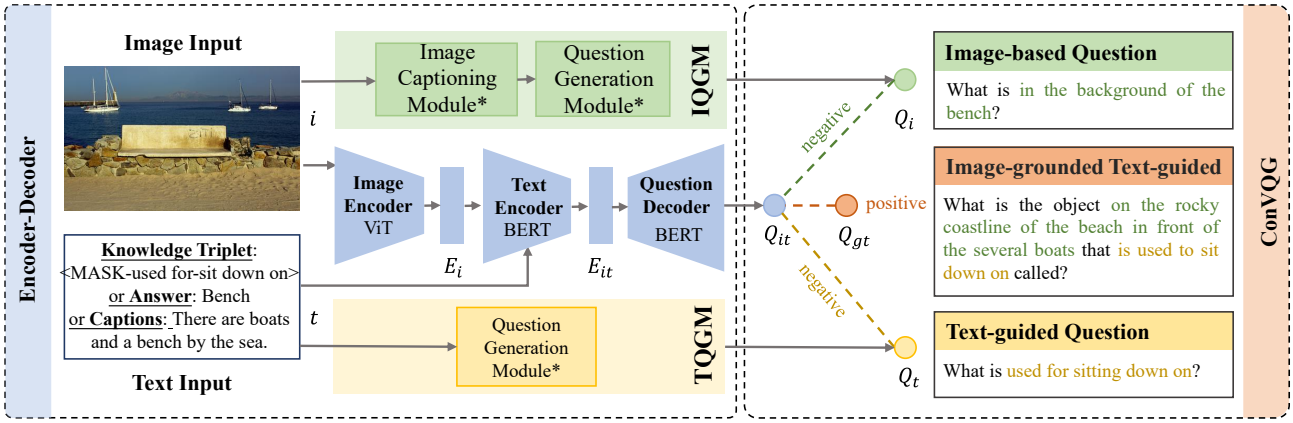


Figure 2: Pipeline of the ConVQG method. During training, an encoder-decoder VQG framework is powered by two additional branches for image-based question generation (IQGM) and text-based question generation (TQGM) (left part, $*$ means the model is frozen). Then, contrastive losses discriminate image-text joint embeddings with the one from single modality only (right part). During inference, only the encoder-decoder framework is activated.

input, splits it into patches, and then feeds them into a transformer encoder (Vaswani et al. 2017) to output a sequence of embeddings E_i: $E_i = \mathbf{ViT}(i)$.

Text Encoder. The text encoder of ConVQG is a variation of the BERT model (Devlin et al. 2018) augmented with additional cross-attention layers at each transformer block to inject visual information into the text encoder. In this way, the text encoder takes as input both the image feature E_i learned by the image encoder and some text t constraining the question to be generated. Such text constraint can take various forms: a knowledge triplet (e.g, *<MASK-used for-sit down on>*),[1] a potential answer (*e.g.*, *bench*), or any other information about the question or the image. They are formulated in natural language t' (shown in supplementary materials). The output of the text encoder is regarded as a joint embedding of the image and text information E_{it}. The text encoder can be formulated as: $E_{it} = \mathbf{BERT}_{encoder}(t', E_i)$.

Question Decoder. The ConVQG question decoder is analogous to the text decoder from BLIP. Essentially, it is a BERT model which replaces the bi-directional self-attention layers with causal self-attention ones. Thus, the inputs to the question decoder are the image-grounded text features learned by the text encoder while the output is the question embedding: $Q_{it} = \mathbf{BERT}_{decoder}(E_{it})$.

Contrastive Learning for VQG

A contrastive learning objective is proposed to generate the question based on both image and text information. The basic idea is that joint embeddings of images and text are supposed to be closer to the embeddings of the question annotations (*i.e.*, the ground truth) while being different from those extracted from unimodal models considering the image (IQGM) or text (TQGM) in isolation.

Image-based Question Generation Module (IQGM). To generate questions based solely on visual information, we first use an image captioning model (**Cap**) from BLIP to generate captions based on the image content. Then we use a question generation model (Ushio, Alva-Manchego, and Camacho-Collados 2022) (**QG**) to generate questions based on these captions. Finally, the generated questions are sent to a sentence-BERT model (Reimers and Gurevych 2019) to obtain the image-based question embeddings Q_i. The models are pre-trained. The IQGM can be denoted as Eq. (1):

$$Q_i = \mathbf{sBERT}(\mathbf{QG}(\mathbf{Cap}(i))). \quad (1)$$

Text-based Question Generation Module (TQGM). The TQGM uses the same pre-trained question generation model (Ushio, Alva-Manchego, and Camacho-Collados 2022) (**QG**) as the IQGM, generating questions from the textual input processed as a sentence (t'). Then, the same sentence-BERT (Reimers and Gurevych 2019) model is used to embed the text-based question:

$$Q_t = \mathbf{sBERT}(\mathbf{QG}(t')). \quad (2)$$

Contrastive Losses for VQG. To ensure VQG focuses both on image and text information, we propose a CL objective. With IQGM and TQGM, we obtain questions that are based only on visual information and text constraints respectively. Then we propose two contrastive losses, one on the image and one on the text. The image contrastive loss CL_{img} enforces the L2-norm between the embedding generated by the IQGM, Q_{it}, and the embedding of the ground truth Q_{gt}, by the same sentence-BERT model, to be closer than the L2-norm between Q_{it} and the image-only question embedding Q_i by a margin m:

$$CL_{img} = \max\left(\|Q_{it} - Q_{gt}\|_2 - \|Q_{it} - Q_i\|_2 + m, 0\right). \quad (3)$$

The text-contrastive loss CL_{txt} is analogous, using the embedding of the text-only model Q_t as negative signal:

$$CL_{txt} = \max\left(\|Q_{it} - Q_{gt}\|_2 - \|Q_{it} - Q_t\|_2 + m, 0\right). \quad (4)$$

Then, the contrastive loss can be formulated as a weighted sum of CL_{txt} and CL_{img} with a parameter α:

$$CL = \alpha CL_{txt} + (1 - \alpha) CL_{img}. \quad (5)$$

[1] Here, the *MASK* token replaces the answer to the question

Finally, the CL loss is combined with a cross-entropy loss CEL between predicted question embeddings and ground truth questions to ensure sufficient information from single modalities. The final loss of the ConVQG model can be represented as:

$$Loss = (\beta CL + CEL)/2, \quad (6)$$

where β is a parameter that can be fixed or tuned; it balances the contributions of the contrastive loss and the cross-entropy loss. In the Results section, we perform experiments to analyse the impact of these hyper-parameters.

Training and inference. IQGM and TQGM are auxiliary, frozen modules. Therefore, the trainable components of ConVQG are only the image and text encoders of the multimodal branch, as well as the text decoder. At inference time, IQGM and TQGM are dropped, and only the multimodal encoder-decoder is used to obtain the question embedding Q_{it}. Then we use beam search, as in the sentence generator from BLIP, to decode the final question from Q_{it}.

Experimental Setup

We compare ConVQG with several methods from the literature, considering different forms of *text inputs*. In this section, we describe the datasets, metrics and the experimental settings that we used for training and evaluation.

Datasets

We evaluate our VQG method on three public datasets: a knowledge-aware benchmark (K-VQG) and two standard VQG benchmarks (VQA 2.0 and VQG COCO).

K-VQG[2] (Uehara and Harada 2023) is a knowledge-aware VQG dataset. It is a large-scale, humanly annotated dataset, where image-grounded questions are tied to structured knowledge (knowledge triplets). Each sample consists of an image, a question, an answer, and a knowledge triplet. K-VQG contains ∼13K images and ∼16K (question, answer) pairs, related to ∼6K knowledge triplets.

VQA 2.0[3] (Goyal et al. 2017) with more than 1M (image, question, answer) triplets, it is the largest and most commonly used dataset for VQG evaluation. Images come from the COCO dataset (Lin et al. 2014), and three (question, answer) pairs were collected per image. In our experiments, we consider two versions of this dataset: VQA 2.0 small (Xu et al. 2020), containing ∼80K images and ∼200K (question, answer) pairs; and VQA 2.0 large (Krishna, Bernstein, and Fei-Fei 2019), which count ∼120K and ∼470K, respectively.

VQG COCO[4] (Mostafazadeh et al. 2016) was created to generate natural and engaging questions for images. It contains 2500 training images, 1250 validation images, and 1250 testing images. Each image contains five natural questions and five ground truth captions. Different from the other two datasets, the answers are not always provided.

[2]https://uehara-mech.github.io/kvqg

[3]https://visualqa.org/download.html

[4]https://www.microsoft.com/en-us/download/details.aspx?id=53670

Evaluation Metrics

Numerical metrics. We use a variety of language generation metrics for evaluation: BLEU (Papineni et al. 2002), 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 (Vedantam, Lawrence Zitnick, and Parikh 2015). Additional information on how these metrics are computed can be found in the supplementary material. Similarly to most work in the literature (Chen et al. 2015; Xie et al. 2021), we use the `pycocoevalcap` package[5] for computing the metrics.

Human evaluation. We use Amazon Mechanical Turk to assess the quality of *model-generated* questions, asking workers to express their preferences about 500 examples extracted from the K-VQG test set. Annotators must choose which of the two questions is better according to two criteria: (1) *grounding to the knowledge triplet*, and (2) *grounding to the image*. They also can indicate when none of the two questions is considered better if they consider that their similarity is too high to make a meaningful choice. Additional details on the sample selection, the human evaluation process, and the instructions and examples given to the workers can be found in the supplementary material.

Experimental Framework

Following BLIP, the image encoder is a ViT-B/16, *i.e.*, a ViT architecture with 12 attention heads, 12 hidden layers, and images divided into 16×16 patches. The text encoder and the question decoder are $\text{BERT}_{\text{base}}$ models, *i.e.*, transformer encoder with 12 attention heads and 12 hidden layers. We initialize the encoder-decoder architecture with the corresponding pre-trained modules from BLIP (Li et al. 2022). Since all BLIP models are publicly available,[6] we choose the "BLIP w/ ViT-B and CapFilt-L" checkpoint for initialization. This model was pre-trained on 129M noisy image-text pairs using CapFilt-L, a captioning and filtering method.

Training was done on six NVIDIA A100-SXM4-40GB with a batch size of 24 each (VQA 2.0 dataset) and four NVIDIA V100-SXM2-32GB with a batch size of 16 each (K-VQG dataset, VQG-COCO dataset). The number of epochs varies depending on the dataset (10 for VQA 2.0, 5 for K-VQG, 5 for VQG-COCO). The starting learning rate is 2e-5 with a weight decay of 0.05.

Results

In this section, we report the VQG results including quantitative, qualitative and human evaluation results. We compare ConVQG with several systems from the literature. For the sake of space, we report here only a subset of results from the literature. Additional results and descriptions of the competing methods can be found in the supplementary material.

[5]https://pypi.org/project/pycocoevalcap/

[6]https://github.com/salesforce/BLIP

Text constraint	*Method*	BLEU-4	METEOR	CIDEr
Answer	IM-VQG	12.37	16.65	0.39
	ConVQG$_{IT}$	**14.30**	**18.67**	**0.78**
Knowledge Triplet	K-VQG	18.84	**22.79**	1.31
	ConVQG$_{IT}$	**20.01**	22.66	**1.53**

Table 1: Results on the K-VQG dataset. The results of IM-VQG are reproduced based on the official code. The results of KVQG are taken from the respective paper.[7]

Test set	*Method*	BLEU-4	METEOR	CIDEr
Small	IVQA	23.9	**35.7**	1.84
	IM-VQG	24.8	26.3	1.94
	iQAN	27.1	26.8	2.09
	Radial-GCN	27.9	27.1	2.10
	MOAG	28.1	27.8	2.39
	ConVQG$_{IT}$	**33.1**	30.0	**2.79**
Large	C3VQG	10.0	13.6	0.47
	IM-VQG	16.3	20.6	0.94
	ConVQG$_{IT}$	**22.4**	**21.8**	**1.78**

Table 2: Results on the VQA 2.0 test sets. The results of the competing methods are taken from the respective papers.[8]

Results on VQG Benchmarks

We train ConVQG on three datasets, with different types of text inputs: knowledge triplets, answers and captions.

Knowledge triplet. Results are reported in Table 1 using the K-VQG dataset (row block *Knowledge Triplet*), with masking the answers as in (Uehara and Harada 2023). ConVQG$_{IT}$ outperforms K-VQG (Uehara and Harada 2023) by 1.17% on BLEU-4 and 0.22 points on CIDEr, and has a slightly lower METEOR score (0.13% difference).

Answer. On the K-VQG dataset, answers can also be used as constraints. In Table 1 (row block *Answer*), ConVQG$_{IT}$ shows an improvement of 1.93% on BLEU-4, 2.02% on METEOR and 0.39 points on CIDEr, with respect to the baseline method. On the VQA 2.0 dataset, samples consist of image, question, answer with no other additional sources of knowledge. Only the answer can be used as a text constraint. Results on the VQA 2.0 dataset, large and small versions, are presented in Table 2. On the VQA 2.0 small, ConVQG$_{IT}$ leads to better performances for all the evaluation metrics. The improvement on CIDEr (0.40 points) demonstrates that the generated questions become semantically similar to ground truth annotations. On VQA 2.0 large, ConVQG$_{IT}$ shows large improvements as well. Indeed, BLEU-4, METEOR, and CIDEr increased by 6.1%, 1.2%, and 0.84 points, respectively, with respect to SOTA approaches.

[8]IM-VQG (Krishna, Bernstein, and Fei-Fei 2019), K-VQG (Uehara and Harada 2023)

[8]IVQA (Liu et al. 2018), IM-VQG (Krishna, Bernstein, and Fei-Fei 2019), iQAN (Li et al. 2018), Radial-GCN (Xu et al. 2020), MOAG (Xie et al. 2021), C3VQG (Uppal et al. 2021)

Method	BLEU-1	METEOR	CIDEr
MDN	36.0	23.4	0.51
MC-BMN	40.7	22.6	0.50
ConVQG$^*_{IT}$	**50.2**	**26.4**	**0.56**

Table 3: Results on VQG-COCO, using captions as text constraint. We report BLEU-1 instead of BLEU-4 to be consistent with the comparison methods. The results for the competing methods are taken from the respective papers.[9]

Caption. On the VQG-COCO dataset, there are no answers nor additional knowledge associated with questions, but captions are used as text inputs. We distinguish ConVQG$^*_{IT}$ from ConVQG$_{IT}$ because when captions are used as text constraints, the captioning step ($\mathbf{Cap}$) is skipped and questions generated by IQGM and TQGM are the same. Results show improvements among all metrics compared with the state-of-the-art methods. Compared with MC-BMN (Patro et al. 2020), BLEU-1, METEOR and CIDEr increase 9.5%, 3.8% and 0.06 points respectively (see Table 3).

Ablation Study

In this section, we perform ablation studies to evaluate the contribution of each of the constrastive objectives. To this end, we distinguish four versions of our ConVQG model:

1. **ConVQG$_B$** is our baseline model, consisting of the multimodal encoder-decoder, without the contrastive modules, trained with cross-entropy loss.
2. **ConVQG$_I$** adds the IQGM module and the image contrastive loss in Eq. (3) to the baseline model.
3. **ConVQG$_T$** adds the TQGM module and the text contrastive loss in Eq. (4) to the baseline model.
4. **ConVQG$_{IT}$** is the full model as shown in Fig. 2, that optimizes the final loss in Eq. (6).

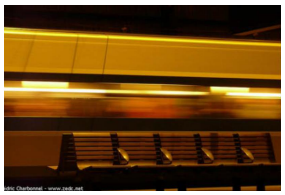
Looking at the performance of ConVQG with some of its components deactivated (Table 4), we see that even the contrastive models using only the image (ConVQG$_I$) or text (ConVQG$_T$) contrastive module outperform the encoder-decoder baseline in all cases but one. For both cases, ConVQG$_{IT}$ works better than ConVQG$_B$, ConVQG$_I$ and ConVQG$_T$, especially for answers as inputs. ConVQG$_{IT}$ outperforms ConVQG$_B$ for 1.35%, 0.89% and 0.14 points on BLEU-4, METEOR and CIDEr respectively.

Parameter Analysis

In the proposed ConVQG method, there are three core parameters: α (Eq. (5)), β (Eq. (6)), both balancing the different parts of the loss and the margin m (Eq. (3) and (4)). We vary their values and test their impact on ConVQG$_{IT}$ on the K-VQG dataset. Results are reported in Table 5.

All in all, these results show that ConVQG is robust to the model hyper-parameters since very small performance variations are observed. Linear β outperforms fixed β values,

[9]MDN (Patro et al. 2018), MC-BMN (Patro et al. 2020)


Text Input	Carrot is a [MASK].	[MASK] is used to sit for a bit.	[MASK] is used for moving across water.
Image-based Question	What is being prepared to be made in the kitchen?	Where is the train stopping at?	What is behind boats near the city of London?
Text-based Question	What is carrot?	What is used to sit for a bit?	What is used for moving across water?
GT Question	What food group does the orange food belong to?	What is the name of the object which people may choose to sit on while waiting for public transportation?	What vehicle type is used to aid people in moving across the water without getting wet?
ConVQG$_{\mathbf{B}}$	What can the orange object that in on the cutting board be used for?	What is the object outside that is used to sit on for a bit?	What is the object in the water that is used for moving across the water?
ConVQG$_{\mathbf{IT}}$	What food group does the orange food on the cutting board belong to?	What is the object on the side of the road that is used to sit on for a bit?	What kind of vehicle placed in the river and is used to moving across water?

Figure 3: Examples from K-VQG dataset with knowledge triplets as inputs. In the text, green color denotes the sequence that is related to image content, while yellow color denotes the information related to the text input. Red color indicates wrong expressions, not related to the image nor the text input. Note: the raw input/output of the model is reported, without correcting grammar or syntax errors made by the generative model.

Text constraint	*Method*	BLEU-4	METEOR	CIDEr
Answer	**ConVQG$_B$**	12.95	17.78	0.64
	ConVQG$_I$	13.95	18.33	0.75
	ConVQG$_T$	13.97	18.03	0.70
	ConVQG$_{IT}$	**14.30**	**18.67**	**0.78**
Knowledge Triplet	**ConVQG$_B$**	18.33	21.47	1.31
	ConVQG$_I$	19.00	21.91	1.38
	ConVQG$_T$	19.11	20.65	1.39
	ConVQG$_{IT}$	**20.01**	**22.66**	**1.53**

Table 4: Ablation studies on K-VQG dataset.

Param.	*Value*	BLUE-4	METEOR	CIDEr
α	0.2	**20.01**	**22.66**	**1.53**
	0.5	19.90	22.60	1.52
	0.8	19.79	22.56	1.52
β	10	19.80	22.55	1.52
	100	19.74	22.39	1.51
	Linear	**20.01**	**22.66**	**1.53**
m	0.2	19.89	**22.66**	**1.53**
	0.5	**20.01**	**22.66**	**1.53**
	0.8	19.68	22.54	1.52

Table 5: Parameter analysis on K-VQG dataset. α from Eq. (5), β from Eq. (6) and m from Eqs. (3) and (4). Linear means β changed linearly during training.[10]

indicating that the contribution of the contrastive loss varies during training. α balances the relative contribution of image contrastive and text contrastive modules, which might vary depending on the dataset and how informative the text constraints are with respect to the image content. For m, metrics are relatively stable, especially for METEOR (max change 0.12%) and CIDEr (max change 0.01 points).

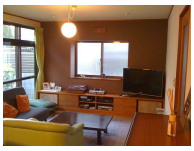
Qualitative Results

Fig. 3 shows generated questions on the K-VQG dataset. For each example, image and text inputs are displayed. Row *Image-based question* corresponds to the question generated by IQGM, while *Text-based question* is the result obtained by TQGM. We compare questions generated by the proposed ConVQG$_{IT}$ with the outputs of the baseline without contrastive learning (ConVQG$_B$) and the annotations.

Comparing the questions generated by the ConVQG versions against the ground truth questions, we observe the following: first, ConVQG$_{IT}$ is able to constrain the question context according to the text inputs more precisely. For example, with the text constraint *Carrot is a [Mask]*, the VQG model is supposed to generate a question about the category or a general description of *Carrot*. The baseline method fails to understand the requirements behind the text input, while the proposed ConVQG$_{IT}$ generated a question that meets the constraint. Second, ConVQG$_{IT}$ provides more information based on both the visual scene (therefore referring to objects in the scene and their relationships) and the text context (formulated as a textual sentence). For instance, in the third example, ConVQG$_{IT}$ replaces *in the water* (ConVQG$_B$) with a more precise generation of the image content (*vehicle placed in the river*). We also provide failure cases, the models sometimes add inappropriate descriptions of images (the middle column) or fail to constrain the question with the text (ConVQG$_B$ in the first column).

The ConVQG model can also be used in inference mode, where a single image and multiple knowledge triplets are

[10] β is increased by a factor of 10 at each epoch, starting from $\beta = 10$.



Input Text: The light bulb.
ConVQG$_{IT}$: What is the name of the appliance that lights up a living room?
Input Text: Shelf is at location of [MASK].
ConVQG$_{IT}$: Where can you find this long wooden object with books?

(a) One image - different text inputs

Input Text: [MASK] is used for floating on water
ConVQG$_{IT}$: What is the vehicle on the water which is used to transportation?

Input Text: [MASK] is used for floating on water
ConVQG$_{IT}$: What is the white object on the beach that is capable of floating on water?

(b) Different images - one text input

Figure 4: Question generation by ConVQG. Given the same image, it can generate different text-guided questions. Given the same text input, it can generate image-specific questions.

given as inputs and vice versa. We show examples of both usages in Figs. 4(a) and 4(b). **In the first case** (*One image - different text inputs*, Fig. 4(a)), the generated questions capture the different constraints provided by the text input. For example, with answer *The light bulb*, the model tries to describe it as *lights up a living room* and *has black and white stripes*. If the text input is changed to *Shelf is at a location of [Mask]*, then the model generates a question about the place and adds more information such as *long wooden object with books*. **In the second case**, if the model is given the same text and different images as inputs (*Different images - one text input*, Fig. 4(b)), ConVQG$_{IT}$ generates image-grounded questions by finding unique image content. In the top example, ConVQG uses the words *vehicle* and *transportation* in the question, showing general understanding provided by the visual cue of people traveling on the boats. In the bottom, the generated question contains the descriptions of the specific boats (*white object*) and of the visual scene (*on the beach*).

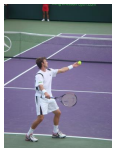
Transfer Results

To demonstrate the generalization ability of ConVQG, we test it in a transfer setting: we train it on the K-VQG dataset and test it on the FVQA (Wang et al. 2017) dataset without further training. FVQA was created for fact-based visual question answering. For each question-answer pair, a fact sentence is provided to clarify the possible commonsense to answer the question, which is used as a text constraint for our transfer settings. Fig. 5 illustrates this experiment. Compared with annotations, the question generated by ConVQG can be grounded to both image and text, which indicates the effectiveness of the contrastive objectives. Quantitative results can be found in supplementary materials.

Human Evaluation Results

In this section, we report the results of the human evaluation performed on Amazon Mechanical Turk, on K-VQG test set.

Input Text: Piano is a instrument.
GT: Whether the instrument in the image is a light or heavy instrument?
ConVQG$_{IT}$: What is the object the woman is holding that is a type of musical instrument?



Input Text: a racket can be used to play tennis.
GT: Which object in this image do people use when playing tennis?
ConVQG$_{IT}$: what is the object the man is holding that is used to hit a ball?

Figure 5: Transfer results on the FVQA dataset.

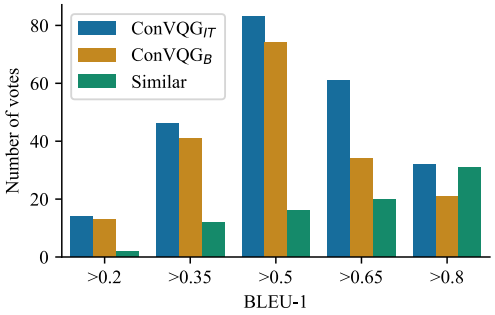


Figure 6: Histogram of human preference by similarity between the two questions, computed using BLEU-1 score.

Among the 500 annotated question pairs, the question generated by ConVQG$_{IT}$ was preferred 236 times; ConVQG$_{B}$ was preferred 183 times; the option "Similar" was chosen 81 times. We compute the similarity between the two questions using the BLEU-1 score. A histogram of the proportion of each of the three choices by degree of similarity between the questions can be found in Fig. 6. The proportion of the "Similar" option chosen by the annotators increases with the similarity between the questions, which is a good way to verify the ability of the workers to correctly tackle the task. Moreover, the contrastive model ConVQG$_{IT}$ is systematically chosen more often than the baseline model, demonstrating the human preference towards the proposed system.

Conclusion

Asking questions in natural language is a fundamental step toward effective visual dialog systems. In this work, we propose contrastive VQG with multimodal guidance from the image content and textual constraints. ConVQG leverages two modality-specific contrastive objectives to guide the content of the question by driving it away from questions generated from single modalities. Our multimodal system allows to control the diversity of questions, and the simultaneous grounding in both modalities. Extensive experiments in standard and knowledge-aware benchmarks show that ConVQG outperforms state-of-the-art methods and has good transfer capacities to unseen datasets. Human evaluation demonstrates that humans prefer ConVQG-generated questions to non-contrastive baselines. These results show that the contrastive objective of ConVQG is key to generating diverse, knowledge-rich, and image-specific questions.

Acknowledgements

We thank the anonymous reviewers for their constructive and thoughtful comments. We also thank Siran Li and Chang Xu for providing codes of baselines; Zeming Chen, Tianqing Fang, Debjit Paul and Valérie Zermatten for providing helpful feedback on earlier versions of this work. We acknowledge the support from the CSC and the EPFL Science Seed Fund. AB also gratefully acknowledges the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Center for Imaging, Sony Group Corporation, and the Allen Institute for AI.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. In *ICCV*, 2425–2433.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *NeurIPS*, 1877–1901.

Chen, F.-L.; Zhang, D.-Z.; Han, M.-L.; Chen, X.-Y.; Shi, J.; Xu, S.; and Xu, B. 2023. VLP: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1): 38–56.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. UNITER: Universal image-text representation learning. In *ECCV*, 104–120.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *CVPR*, 326–335.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *WMT*, 376–380.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Gan, Z.; Li, L.; Li, C.; Wang, L.; Liu, Z.; Gao, J.; et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4): 163–352.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12): 3618–3623.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 6904–6913.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*, 18661–18673.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Klein, T.; and Nabi, M. 2021. Attention-based contrastive learning for winograd schemas. In *EMNLP-Findings*, 2428–2434.

Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2019. Information maximizing visual question generation. In *CVPR*, 2008–2018.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, Y.; Duan, N.; Zhou, B.; Chu, X.; Ouyang, W.; Wang, X.; and Zhou, M. 2018. Visual question generation as dual task of visual question answering. In *CVPR*, 6116–6124.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*, 740–755.

Liu, F.; Xiang, T.; Hospedales, T. M.; Yang, W.; and Sun, C. 2018. Inverse visual question answering: A new benchmark and VQA diagnosis tool. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 460–474.

Mostafazadeh, N.; Misra, I.; Devlin, J.; Mitchell, M.; He, X.; and Vanderwende, L. 2016. Generating natural questions about an image. In *ACL*, 1802–1813.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2023. GPT-4 Technical report. Technical report.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 311–318.

Patro, B.; Kurmi, V.; Kumar, S.; and Namboodiri, V. 2020. Deep bayesian network for visual question generation. In *WACV*, 1566–1576.

Patro, B. N.; Kumar, S.; Kurmi, V. K.; and Namboodiri, V. P. 2018. Multimodal differential network for visual question generation. In *EMNLP*, 4002–4012.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP*, 3982–3992.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NeurIPS*.

Saeed, A.; Grangier, D.; and Zeghidour, N. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP*, 3875–3879.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Uehara, K.; and Harada, T. 2023. K-VQG: Knowledge-aware visual question generation for common-sense acquisition. In *WACV*, 4401–4409.

Uppal, S.; Madan, A.; Bhagat, S.; Yu, Y.; and Shah, R. R. 2021. C3VQG: Category consistent cyclic visual question generation. In *ACM MM Asia*.

Ushio, A.; Alva-Manchego, F.; and Camacho-Collados, J. 2022. Generative language models for paragraph-level question generation. In *EMNLP*, 670–688.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Vedd, N.; Wang, Z.; Rei, M.; Miao, Y.; and Specia, L. 2022. Guiding visual question generation. In *ACL*, 1640–1654.

Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017. FVQA: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10): 2413–2427.

Xie, J.; Cai, Y.; Huang, Q.; and Wang, T. 2021. Multiple objects-aware visual question generation. In *ACM MM*, 4546–4554.

Xie, J.; Fang, W.; Cai, Y.; Huang, Q.; and Li, Q. 2022. Knowledge-based visual question generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7547–7558.

Xu, X.; Wang, T.; Yang, Y.; Hanjalic, A.; and Shen, H. T. 2020. Radial graph convolutional network for visual question generation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4): 1654–1667.

Zhang, S.; Qu, L.; You, S.; Yang, Z.; and Zhang, J. 2017. Automatic generation of grounded visual questions. In *IJCAI*, 4235–4243.