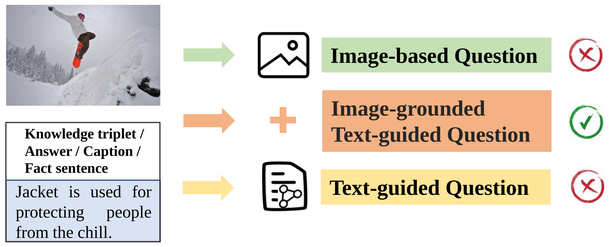
ConVQG: Contrastive Visual Question Generation with Multimodal Guidance

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

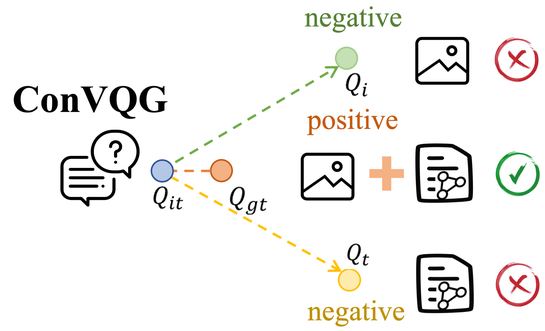
EPFL, Switzerland

1. Motivation

- **VQG**: Generating questions using **textual constraints** while enforcing a high relevance to the **image content**.
- VQG systems often **ignore** one or both forms of grounding.



ConVQG at a glance



- Driving the joint embedding away from single modality questions by **contrastive objectives**.
- Generating text-guided, image-grounded and knowledge-enriched questions for images.

2. Method

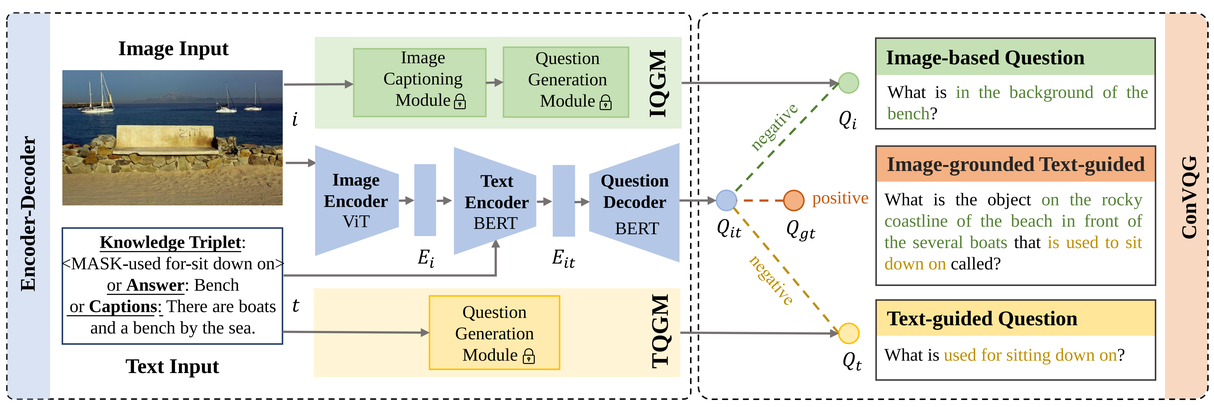


Image Contrastive

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

Text Contrastive

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

Final Objectives

Contrastive Loss: $CL = \alpha CL_{txt} + (1-\alpha)CL_{img}$.

Cross-entropy Loss: CEL for question generation

Total Loss: $Loss = (\beta CL + CEL)/2$.

3. Question Generation Results

Grounding

→ ConVQG generations have **a better grounding to both image and text** compared to the baseline.

Text Input: Carrot is a [MASK].
ConVQG$_B$: What can the orange object that in on the cutting board be used for?
ConVQG$_{IT}$: What food group does the orange food on the cutting board belong to?

Text Input: [MASK] is used for moving across water.
ConVQG$_B$: What is the object in the water that is used for moving across the water?
ConVQG$_{IT}$: What kind of vehicle placed in the river and is used to moving across water?

Diversity

→ ConVQG allows to control the **diversity of questions** with different text/image 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?

Transferability

→ ConVQG can transfer to **unseen text constraints** (fact sentences) without further training.

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: boat is related to fishing.
GT: What do you need in this image if you want to fishing?
ConVQG$_{IT}$: what is the object on the water that is used for fishing?

4. Evaluation Results

- **Benchmark Comparisons**

→ Standard and knowledge-aware benchmarks.
→ Different types of text constraints.

1) **Knowledge triplets** as constraints

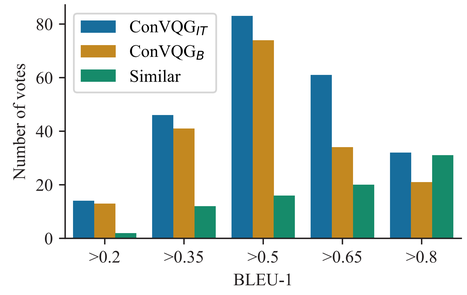
Dataset	*Method*	BLEU-4	CIDEr
K-VQG	K-VQG [1]	18.8	1.31
	ConVQG$_B$	18.3	1.31
	ConVQG$_{IT}$	**20.0**	**1.53**

2) **Answers** as text constraints

Dataset	*Method*	BLEU-4	CIDEr
K-VQG	IM-VQG [2]	12.4	0.39
	ConVQG$_B$	13.0	0.64
	ConVQG$_{IT}$	**14.3**	**0.78**
VQA2.0 Small	MOAG [3]	28.1	2.39
	ConVQG$_{IT}$	**33.1**	**2.79**
VQA2.0 Large	IM-VQG [2]	16.3	0.94
	ConVQG$_{IT}$	**22.4**	**1.78**

- **Human Evaluation**

→ Pairwise comparison among 500 examples.
→ Human preference on contrastive generations.



References and Materials

[1] Kohei Uehara and Tatsuya Harada. K-VQG: Knowledge-aware visual question generation for common-sense acquisition. In *WACV*, 2023.

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

[3] Jiayuan Xie, Yi Cai, Qingbao Huang, and Tao Wang. Multiple objects-aware visual question generation. In *ACM MM*, 2021.

Questions? Contact us!

Page: https://limirs.github.io/ConVQG/
Email: li.mi@epfl.ch