Knowledge-Guided Cross-Topic Visual Question Generation

Hongfei Liu[1,2], Guohua Wang[1,2], Jiayuan Xie[3], Jiali Chen[1,2], Wenhao Fang[1,2], Yi Cai[1,2,*]

Speaker：Hongfei Liu

1.School of Software Engineering, South China University of Technology
2.Key Laboratory of Big Data and Intelligent Robot (South China University of Technology) Ministry of Education
3.Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China

Contents

1 Introduction

Visual Question Generation

Input Image

Constraints

Answer: white

Model

VQG Model

Question

What color is the vest of the person on the left?

Question Type: What

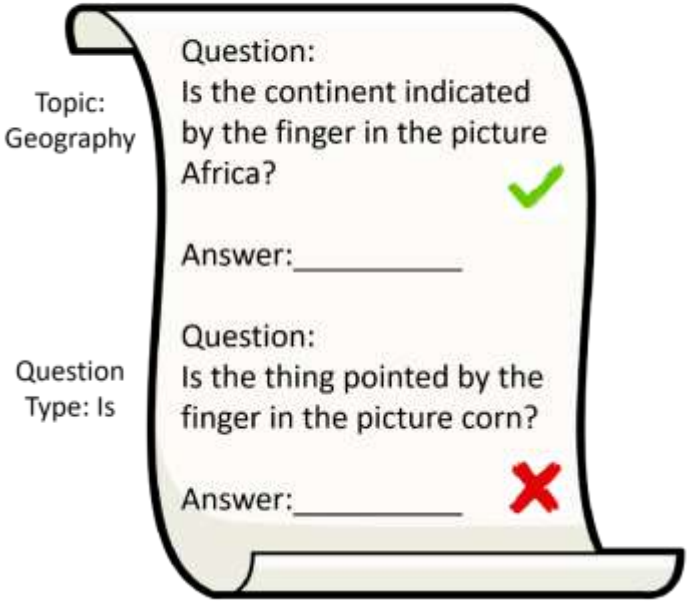
Constraints

Application

Use Topic as Constraints

Chat About Basketball

Geography Exam



Cross-Topic Learning VQG

South China University of Technology

Existing datasets generally contain only a small selection of common topics, and existing models struggle to generate high-quality questions in cross-topic scenarios.

In this paper, we introduce the Cross-Topic Learning Visual Question Generation (CTL-VQG) task, which is designed to generate questions across a range of topics, using a limited set of labeled data and extending effectively to those unlabeled topics, especially situations that are uncommon or require expertise to label.

To address the CTL-VQG task, a VQG model must learn valid paradigms from the data associated with annotated topics. According to Duan et al. (2008), questions can be divided into two components: structural and content information. Specifically, structural information is influenced by the type of question, such as “which” or “is”, whereas content information predominantly pertains to the topic.

Figure 2(a) illustrates the distribution of all questions under the topic “color”, with an even distribution across different question types. Figure 2(b) demonstrates that when specific content, such as the object "dog" associated with the topic "color" is removed, the questions of each type cluster together. This suggests that cross-topic learning primarily captures content relevant to new, unannotated topics while preserving the structural integrity of the questions.

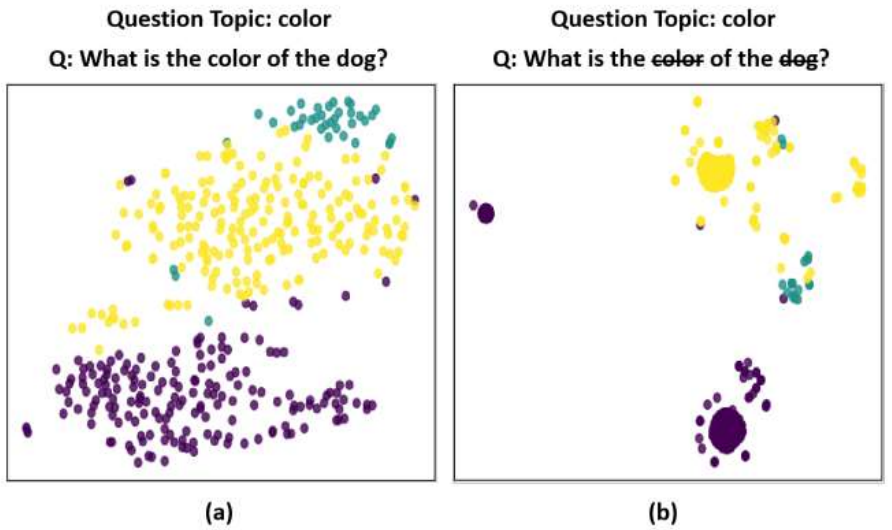


Figure 2: Question embedding distribution in different question types with the question topic “color”. Each point represents the coordinates of a question embedding after dimensionality reduction and normalization. The points are color-coded to represent questions of different question types. Figure (a) represents the original question embedding. Figure (b) represents the embedding of the question without topic-related content.

2 Methodology

KC-VQG Model

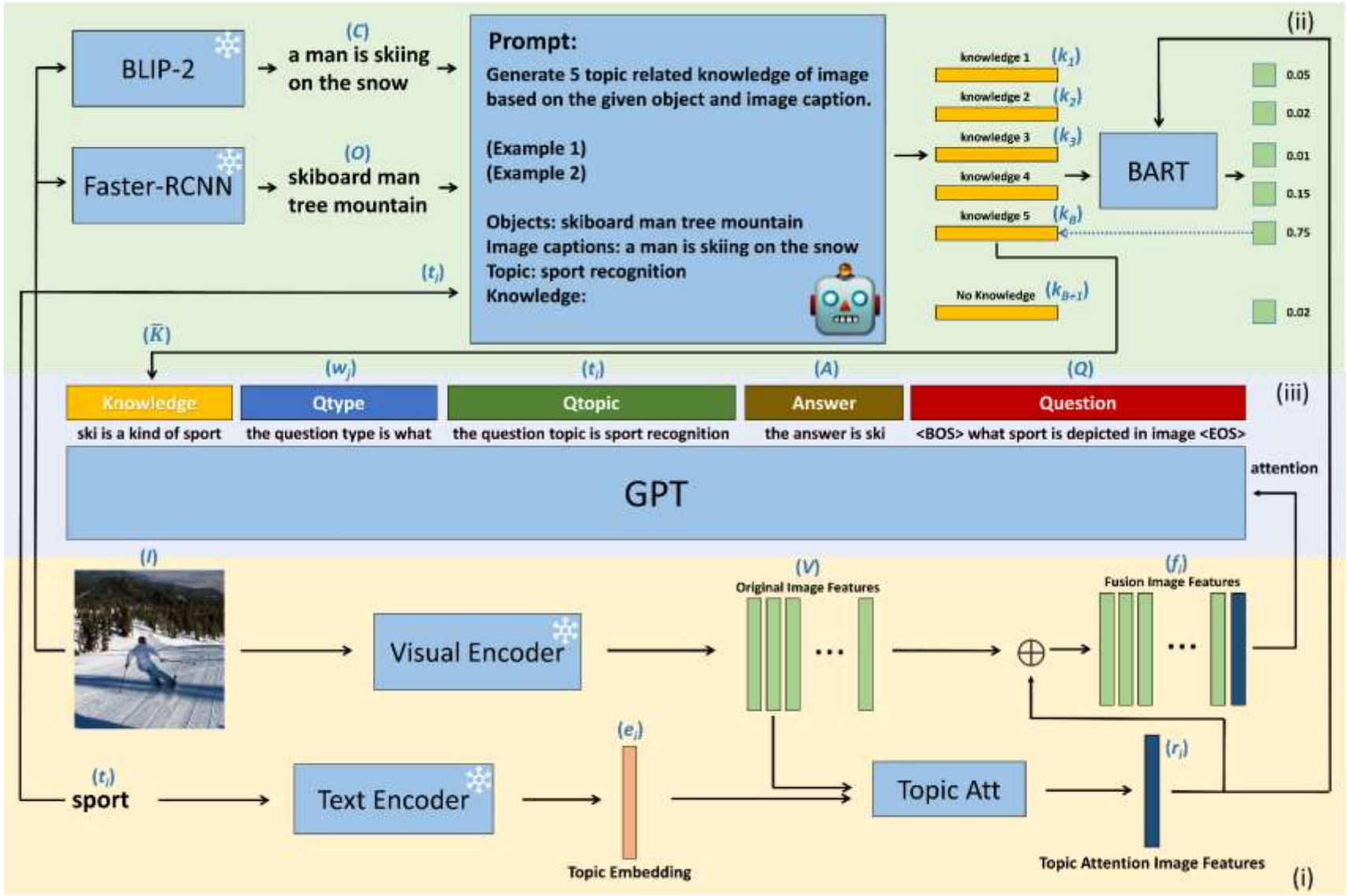


Image-Topic Knowledge Extractor

Topic and Type-Guided Question Decoder

Image-Topic Feature Extractor

3 Experiment

Experiment Settings

South China University of Technology

Dataset: TDIUC (Kaffe and Kanan, 2017) , 11 topics
1 topic in testing set, the other 10 topic in training and evaluation set

Baseline models:
IMVQG (Krishna et al., 2019) – seq2seq based model
VQG-GCN (Xu et al., 2021) – GCN based model
MOAG (Xie et al., 2021) – GCN based model
ClipCap (Mokady et al., 2021) – transformer based model
KB-VQG (Xie et al., 2022) – knowledge based model
GPT-3.5 (OpenAI, 2023) – LLM

Ablation models:
KC-VQG w/o KD: KC-VQG model without using knowledge discriminator
KC-VQG w/o TA: KC-VQG model without using fusion image features in decoder

Automation evaluation metrics:
BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), BERT-score (Zhang et al., 2020)

Human evaluation criteria:
Fluency (Flu), Image relevance (Img_rl), Answer relevance (Ans_rl), Topic relevance (Top_rl)

Automatic Evaluation Results

South China University of Technology

Model	BLEU-4	METEOR	BertScore
IMVQG	39.56	36.50	94.49
VQG-GCN	25.11	34.48	94.90
MOAG	19.66	31.56	88.34
ClipCap	33.61	32.20	92.68
KB-VQG	39.90	35.49	94.71
GPT-3.5	3.81	11.64	90.04
KC-VQG w/o KD	46.34	39.14	95.99
KC-VQG w/o TA	**47.82**	39.06	**96.31**
KC-VQG(Ours)	46.85	**39.20**	95.72

Table 2: The automatic evaluation result of non-cross-topic scenario. **Bold**: the maximum value in the column.

Topic	Model	BLEU-3	BLEU-4	METEOR	ROUGE$_L$	CIDEr	BertScore
activity recognition	IMVQG (Krishna et al., 2019)	16.55	6.72	18.06	43.76	28.54	86.00
	VQG-GCN (Xu et al., 2021)	1.25	0.00	11.12	27.13	5.10	85.79
	MOAG (Xie et al., 2021)	6.70	0.00	14.13	51.08	4.85	86.34
	ClipCap (Mokady et al., 2021)	8.27	4.99	15.98	32.30	16.41	85.87
	KB-VQG (Xie et al., 2022)	15.63	6.43	16.75	41.74	41.25	86.76
	GPT-3.5 (OpenAI, 2023)	7.02	5.47	10.32	32.16	**86.70**	91.35
	KC-VQG w/o KD	12.36	9.36	12.37	39.27	53.07	88.51
	KC-VQG w/o TA	17.79	11.40	15.57	48.50	34.41	88.72
	KC-VQG(Ours)	**27.65**	**17.24**	**20.90**	**53.93**	50.45	**92.33**
attribute	IMVQG (Krishna et al., 2019)	8.11	2.41	11.78	34.80	1.96	85.15
	VQG-GCN (Xu et al., 2021)	1.60	0.00	11.86	28.90	2.35	87.81
	MOAG (Xie et al., 2021)	17.66	4.39	11.94	47.67	3.63	82.39
	ClipCap (Mokady et al., 2021)	9.21	3.58	13.35	32.25	9.56	84.48
	KB-VQG (Xie et al., 2022)	10.04	0.00	12.05	33.67	7.82	85.77
	GPT-3.5 (OpenAI, 2023)	5.74	2.26	9.57	30.80	21.49	87.78
	KC-VQG w/o KD	18.49	11.53	15.84	48.30	57.39	88.41
	KC-VQG w/o TA	18.19	11.24	15.20	50.99	48.03	88.99
	KC-VQG(Ours)	**26.92**	**16.63**	**17.19**	**51.82**	**72.44**	**89.89**
color	IMVQG (Krishna et al., 2019)	10.61	0.00	12.48	47.06	6.23	86.87
	VQG-GCN (Xu et al., 2021)	2.88	0.92	11.16	28.73	4.82	87.84
	MOAG (Xie et al., 2021)	44.15	36.19	26.56	62.55	6.51	88.57
	ClipCap (Mokady et al., 2021)	13.80	9.74	20.43	38.43	10.06	85.76
	KB-VQG (Xie et al., 2022)	9.99	0.00	12.86	46.73	17.51	87.16
	GPT-3.5 (OpenAI, 2023)	6.63	4.59	13.10	33.83	34.85	90.90
	KC-VQG w/o KD	41.83	34.59	30.22	62.70	72.17	91.22
	KC-VQG w/o TA	43.49	36.10	**31.10**	**65.10**	**110.21**	**92.61**
	KC-VQG(Ours)	**44.53**	**37.24**	30.62	64.32	93.79	91.62
counting	IMVQG (Krishna et al., 2019)	0.00	0.00	3.82	5.31	5.44	84.48
	VQG-GCN (Xu et al., 2021)	5.03	1.01	7.22	10.18	9.98	87.32
	MOAG (Xie et al., 2021)	0.00	0.00	4.54	4.28	2.00	79.96
	ClipCap (Mokady et al., 2021)	6.16	3.73	17.25	24.47	9.73	85.02
	KB-VQG (Xie et al., 2022)	0.72	0.30	4.10	5.59	4.28	82.66
	GPT-3.5 (OpenAI, 2023)	13.37	9.40	19.64	40.81	65.07	92.41
	KC-VQG w/o KD	35.10	28.01	33.10	59.07	197.79	93.35
	KC-VQG w/o TA	31.25	25.73	29.20	50.41	174.96	92.34
	KC-VQG(Ours)	**37.33**	**30.74**	**33.97**	**59.21**	**208.59**	**93.63**
object recognition	IMVQG (Krishna et al., 2019)	0.98	0.00	11.26	48.82	4.35	84.25
	VQG-GCN (Xu et al., 2021)	1.60	0.00	10.29	29.24	0.75	85.81
	MOAG (Xie et al., 2021)	0.00	0.00	10.70	45.05	4.56	79.12
	ClipCap (Mokady et al., 2021)	6.36	1.75	13.59	32.92	2.09	83.75
	KB-VQG (Xie et al., 2022)	1.53	0.00	10.77	41.52	9.76	84.90
	GPT-3.5 (OpenAI, 2023)	2.20	0.87	8.82	27.90	3.38	88.80
	KC-VQG w/o KD	6.78	3.06	11.81	46.58	9.38	85.39
	KC-VQG w/o TA	6.12	1.48	12.25	47.33	2.68	85.82
	KC-VQG(Ours)	**23.88**	**16.29**	**19.28**	**57.94**	**63.20**	**89.41**

Human and Inference Speed Evaluation Results

South China University of Technology

Model	Flu	Img_rl	Ans_rl	Top_rl
IMVQG	0.39	0.35	0.46	0.38
VQG-GCN	2.58	0.48	0.23	0.64
ClipCap	1.89	1.76	1.20	2.02
MOAG	1.25	0.89	0.35	1.01
KB-VQG	0.93	1.88	0.72	2.13
GPT-3.5	**4.22**	2.33	**3.84**	3.25
KC-VQG	3.84	**3.12**	2.44	**3.38**

Table 3: The human evaluation results in all cross-topic learning scenarios. **Bold**: the maximum value in the column.

Model	model_cost	api_cost	total
IMVQG	38.36	-	38.36
VQG-GCN	228.81	-	228.81
ClipCap	532.89	-	532.89
MOAG	**4.72**	-	**4.72**
KB-VQG	716.28	-	716.28
KC-VQG	115.33	126.92	242.25

Table 4: The inference speed evaluation results (seconds/1000 questions). The api_cost column in the table indicates the time taken by our KC-VQG model in the Knowledge Generator module to generate knowledge. **Bold**: the minimum value in the column.

4 Case Study

04 Case Study

Topic: **attribute**

Generated Knowledge:

1. Glasses can be used to correct vision or protect the eyes from sunlight.
2. Trees are living organisms that provide shade, oxygen, and habitats for animals.
3. A vest is a sleeveless garment that … .
4. A face is the front part of a person's head where … .
5. A tie is a long, thin piece of fabric that is worn around the neck in a knot, usually as a decorative accessory.
6. No Knowledge

Generated Questions:

KC-VQG (ours)	: What material is the tie shown in the picture? ✓
IMVQG	: What is to man doing doing the?
VQG-GCN	: What is the man doing?
CLIPCap	: What is he wearing what is he doing at the office?
MOAG	: what is behind behind the?
KB-VQG	: How is the man doing doing the?
Ground Truth	: What is the vest made of?

(a)

Topic: **counting**

Generated Knowledge:

1. There is one girl in the image.
2. The girl is wearing glasses.
3. There is one tree in the background.
4. The girl is wearing one pink dress.
5. The girl is playing with one frisbee.
6. No Knowledge

Generated Questions:

KC-VQG (ours)	: How many cars are in the photo? ✓
IMVQG	: What is is is the?
VQG-GCN	: What are the people doing?
CLIPCap	: How is her wearing around her neck most important?
MOAG	: What is the the the?
KB-VQG	: What is is is the any?
Ground Truth	: How many cars are there?

(b)

Figure 4: Case study of generated questions by our model and baseline models. The knowledge highlighted in red in the figure represents the most suitable knowledge chosen by the knowledge discriminator.

華南理工大學
South China University of Technology

Thank you for your listening!

Speaker：Hongfei Liu