ChatVQG: A VQG dataset containing diversified conversational questions and guiding information for proactive chatting with the elderly

Yuyang Tao
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
taoyuyang33@bistu.edu.cn

Yuru Jiang
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
yurujiang@126.com

Mengyuan Li
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
limengyuan0114@gmail.com

Yangsen Zhang
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
zhangyangsen@163.com

***Abstract*—With the increasing global aging problem, the application of NLP technology to elderly care services has become a crucial field o f r esearch. C urrently, c are c hatbots u tilized in hospitals, nursing homes, and other institutions can only passively provide services when elderly individuals express their needs actively. We propose that by combining VQG technology with care chatbots, they can provide proactive chat services to the elderly. However, existing VQG datasets are not designed for chat scenarios specific t o c hatting-with-elderly. T o a ddress t his, we developed a VQG dataset called ChatVQG. It includes multiple diverse questions for each picture, along with two guiding labels per question: question topic and question object, which assist in instructing the generation model. We proposed a prompt-based text-to-text method of guided question generation and a self-guided pipeline to implement proactive questioning. We used BLIP-2 as the baseline model and fine-tuned i t o n ChatVQG. The experimental results demonstrate the effectiveness of our proposed method and pipeline.**

***Index Terms*—Active Questioning, Elderly Care, Visual Question Generation, Guiding Information, Image-text Generation, Multi-modal**

I. Introduction

The world's aging population is growing. In 2019, 102 countries and regions (50.25% of the total) were considered aging societies, and by 2050, this number is expected to rise to 158 (77.83%). Elderly people need more attention and help, especially when their families are not around. However, there aren't enough caregivers to meet this demand. Also, as technology advances, many elderly people are feeling left out, which makes them lonely. Loneliness can harm their mental health, leading to depression, anxiety, and a higher risk of dementia.

Measures should be taken to alleviate loneliness, improve social participation, and enhance the quality of life for the elderly. While family and friends' company is ideal, it's often not possible due to loved ones passing away or children's work commitments, leaving elderly individuals alone. Consequently, new technologies like chatbots and other assistive tools offer a viable alternative by providing entertainment, memory support, medication reminders, and enabling communication with family, friends, or healthcare professionals. This can boost the elderly's positivity and combat loneliness [1].

Elderly people often have trouble starting conversations because they lack motivation, leading to long periods of inactivity. We suggest using chatbots along with Visual Question Generation (VQG) technology to help with this. As shown in Fig. 1, the chatbot uses a camera to take pictures of everyday scenes or access a collection of photos. Then, it shows these pictures to the elderly and asks them questions about what they see. This encourages conversation and provides a chance for the elderly to practice their language skills, keep their minds active, combat loneliness, and improve their mental well-being. Ultimately, this can make their lives better and offer a new solution for our aging society.

When questioning the elderly, it's important to follow a few guidelines. First, maintain a natural and engaging tone in your questions. Second, ensure the questions are clear and avoid using overly specialized or complex language. Third, ask about topics and subjects that pique the interest of the elderly. Lastly, always approach the questioning with respect and politeness, treating them as you would a respected elder.

We created a new dataset called ChatVQG to address the limitations of existing datasets for elderly chatting scenarios. To make the annotation process more efficient, we used ChatGPT for pre-annotation in a two-stage annotation process. We divided the visual question generation task into three subtasks and unified them into a visual-text generation task. We developed a self-guided pipeline, allowing the model to use

979-8-3503-3078-6/23/$31.00

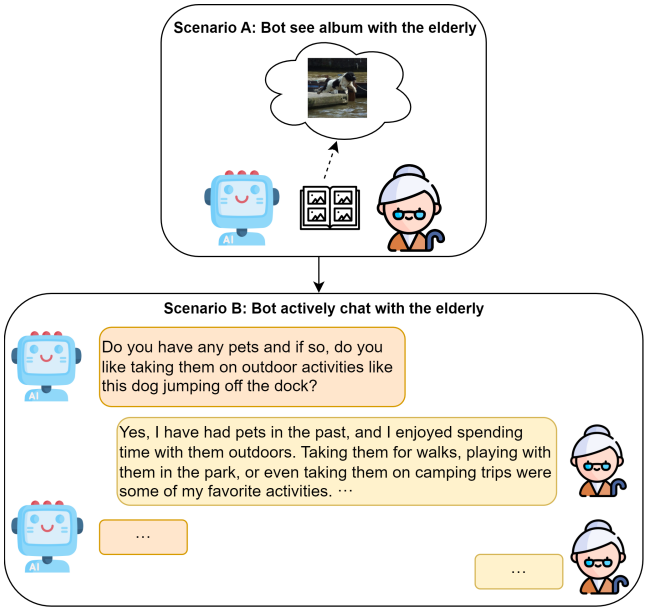

Fig. 1. A scenario of elderly-health-care chatbot raising an engaging question about a photo to elderly for chatting.

image input to choose a question topic and generate related questions. Our baseline model is BLIP-2 Flan T5.

Our contributions in this paper are as follows:

(1) **ChatVQG:** We propose a VQG dataset for chatting-with-elderly cenario, called ChatVQG, which consists of 1742 images with a total of 5358 appropriate questions for asking to elderly people, along with corresponding topic and object labels.

(2) **A Prompt Based Text-to-Text Method for Using and Training on ChatVQG**

(3) **A Self-Guided Question Generation Pipeline**

(4) **Experiments and Evaluations that show the effectiveness of proposed dataset and methods**

II. Related Work

Visual Question Generation: The task of visual question generation is to explore the connection between vision and language. Mostafazadeh [2] introduced the VQG task and its first dataset, which focused on creating engaging questions people typically ask when sharing photos on social media. They wanted systems to generate questions that would encourage responses. Another dataset, MVQG [3], includes multiple images to enhance the model's applicability. However, these studies prioritize engaging the majority and may not suit everyone. Our work, in contrast, targets elderly individuals, aiming to generate questions that align with their characteristics and stimulate conversations.

Image Caption: The image captioning task involves describing an image's content using text, which is crucial for generating questions based on detailed image information. Early image captioning methods relied on retrieval techniques [4], [5], [6], [7], but they struggled to produce image-specific and semantically accurate captions. Newer methods operate in a multimodal space, where they analyze the image's visual content and then use a language model to generate captions [8], [9], [10], [11]. These modern methods are more capable of producing novel captions with higher semantic accuracy. Recent research has shown that multimodal large models excel in image captioning tasks. BLIP [12] introduced the CapFit method, which generates higher-quality textual descriptions for images. BLIP-2 [13] set a new zero-shot captioning benchmark. ChatCaptioner [14] leverages ChatGPT to ask image-related questions to BLIP-2, encouraging it to provide more visual information. By continually extracting new visual details from BLIP-2's responses, ChatCaptioner generates more comprehensive image descriptions.

Visual and Language Models: In recent years, researchers have been studying models that combine visuals and words. In our work, we need a model that not only does the job but also knows a lot about common sense. But because of limited data, models like BERT [15] can't learn much common sense. For instance, extreme sports like diving have specific knowledge that normal models can't learn, making it hard to generate good questions. So, we need big and powerful language models. Some popular ones are VL-BERT [16], VLT5 [17], BLIP, and BLIP-2. VL-BERT uses two encoders with a shared layer to mix visuals and text. VLT5 can handle text, images, and videos using a shared technique. BLIP is good for various tasks and BLIP-2 is cost-effective by building on existing models. BLIP-2 does really well, like generating text from images with just a description.

III. ChatVQG Dataset

A. Dataset Construction

We use images already chosen by VQG dataset [2], as a first filter. Then, we remove some images that aren't good for asking questions to elderly folks. A flowchart of the dataset construction process is shown in Figure 2.

(1) **Image Source:** We selected 5000 images from the MS COCO image dataset [18] that are used in the $\text{VQG}_{\text{coco-5000}}$ dataset [2] as our initial labeling objects. We further discarded unsuitable or content-repeated images during the labeling process, and finally 1742 pictures remained.

(2) **Guiding label design:** Images contain rich information, and there are multiple perspectives and goals for questioning in each image. In order to generate diverse questions in a controllable way, we designed two guiding labels: question topic and question object. Each question corresponds to a set of guiding labels for topic and object.

In order to determine the scope of question topic, we prepared 7 common topics in everyday conversation and designed a questionnaire. We invited 12 elderly people over 60 years old to vote for their favorite topics for chatting. We asked the elderly to choose the five most preferred topics out of the seven. The top five topics with the most votes are: 1) **travel experiences**, 2) **health and lifestyle**, 3) **hobbies and interests**, 4) **family and children**, and 5) **history and culture**. We used these five topics as the selection range for our

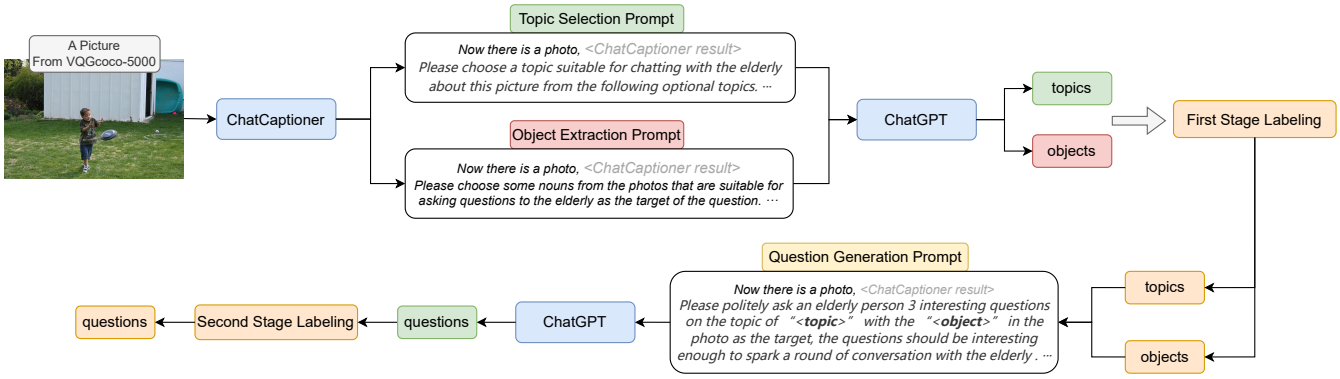

Fig. 2. ChatVQG dataset construction process.

question topic labels. For question object, we selected suitable scenes or objects in the images for questioning based on the requirements of the chatting-with-elderly scene.

(3) **Labeling Process:** The labeling process consists of two stages: (a) **Topic Selection & Object Extraction** and (b) **Question Writing**.

We employed ChatGPT as a supplementary tool in our workflow. At the start of each phase, we initially employed ChatGPT for pre-annotation. Subsequently, we reviewed, supplemented, or manually adjusted the pre-annotated outcomes.

In the first stage, we selected the question topics and objects that can be asked about the corresponding image. We first corrected factual errors in the image detail generated by ChatCaptioner, such as descriptive errors related to color, orientation, and quantity. Then, based on the preliminary results of topics and objects generated by ChatGPT, we manually screened and labeled the appropriate topics and objects. For images with few useful details, excessive repetition, or containing sensitive information, they would be discarded.

In the second stage, We used ChatGPT to generate three sample questions for each combination of topic and object. We either chose the most suitable one from the three samples, or write a question of our own that is highly relevant to the topic and object. If the combination of the topic and object is unsuitable the question would be discarded.

B. Data Analysis

To demonstrate the characteristics of ChatVQG, we collected various data as shown in Table I. A total of 1742 images from COCO were selected, and 5358 questions were written by two of our colleague, averaging 3.07 diversified questions per image. A sample of ChatVQG is shown in Figure 3.

We also analyzed the top-15 most frequently occurring 3-gram in our questions, as presented in Table II. It can be observed that our questioning approach primarily focuses on seeking opinions and preferences of the elderly, such as "do you think", "do you like", and discussing their past experiences, for instance, "have you ever", "you ever tried", "you ever been", showing the characteristics of the chatting-with-elderly scenarios. Additionally, we calculated the word

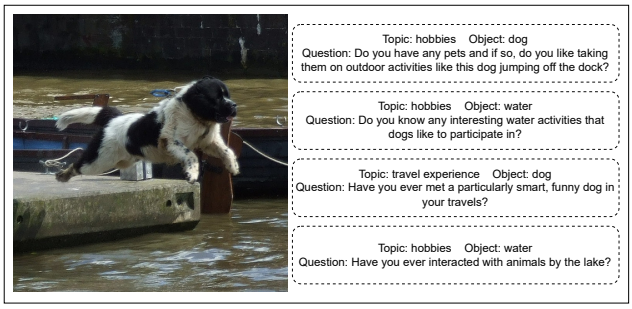

Fig. 3. A sample from ChatVQG.

count in the questions, as shown in Table I. A total of 4464 words appeared in our generated questions. Among these, the nine most frequent words are shown in Table III. The most frequent word is "you", reflecting the nature of our dataset, which places emphasis on caring for and accompanying the elderly, rather than the content of the images. Furthermore, we examined the fifteen most selected question objects, as depicted in Table III. These question objects mainly comprise common objects such as pets and scenery, indicating that our generated questions aim to be easily understandable by the general public. The average length of our questions is 13.48, indicating that our questions possess abundant information and a certain level of depth.

TABLE I
CHATVQG STATS

all images	1742
all questions	5358
average questions per image	3.076
unique objects	1127
unique words	4464
average question length	13.48

C. Application Scenarios

In terms of real-life elderly care scenarios, ChatVQG can be used to train a visual question generation model to ask polite

TABLE II
TOP 15 FREQUENT 3-GRAM IN OUR QUESTIONS

have you ever	do you think	do you have
you have any	do you like	what do you
you ever tried	when you were	you ever been
you like to	how do you	you ever traveled
what kind of	do you usually	you think the

TABLE III
TOP 9 FREQUENT WORDS AND TOP 15 FREQUENT SELECTED OBJECTS

you	do	have
ever	your	what
think	like	and
cat	dog	kitchen
beach	horse	bus
field	pizza	train
laptop	living room	couch
vase	frisbee	cake

and topical conversation questions to the elderly based on given photos. Such VQG model can enable healthcare robots to proactively ask questions.

ChatVQG can also be used for research on question generation technology. QG's open-ended nature makes the focus of question generation technology not on accuracy, but on controllability and diversity. ChatVQG can be used as a training or evaluation dataset for both controllable QG techniques and diverse QG techniques.

IV. METHOD

In this section, we will introduce an effective method that fully leverages ChatVQG to train an image-text generation model. We chose BLIP-2 [13] as the baseline model for ChatVQG and selected the model that uses Flan T5 [19] as the LLM in the BLIP-2 series because Flan T5 has received a lot of instruction tuning and chain-of-thought training, giving it excellent instruction understanding and reasoning capabilities, which is very important for controllable question generation.

A. Task Definitions

We designed three training tasks to enable the model to self-guide and generate appropriate questions in real-world application scenarios by only using images, without requiring additional annotations.

(1) **Topic Selection:** Given a fixed topic selection range T, the model needs to select one or more appropriate question topics $t \in T$ based on the image content.

(2) **Object Extraction:** According to the image content, select items or scenes suitable for questioning as the question object o.

(3) **Topic-and-Object-Aware Question Generation:** Based on the image content, generate a question q consistent with the given question topic t and question object o.

B. Prompt Based Text-to-Text Method

We follow the text-to-text idea proposed by T5 [20] and convert all tasks into text generation tasks. We design input prompt templates and output templates for each task, aiming to convert the input and output data of different tasks into smooth natural language text, so that the model can be naturally trained on different tasks with the same loss function. Given an input image $\mathbf{p}$, an input text $\mathbf{x}$, and an output text $\mathbf{y}$ of length N, the model's loss function for generating output text based on the image and input text is shown in equation 1.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} logP(y_i|y_1 \dots y_{i-1}, \mathbf{x}, \mathbf{p}) \quad (1)$$

C. Self-Guided Generation Pipeline

We propose a self-guided question generation pipeline so that the model can generate diverse topical conversational questions without any help of any other system or human effort, as shown in Fig. 4. It requires only an image. After being trained on the topic selection task and the target extraction task, the model has the ability to select appropriate question topics and question targets on its own, so the entire pipeline is divided into two stages. First, the model will determine the appropriate question topic and extract the relevant question objects from the image. Then the model will generate questions related to each combination of question topic and question object.

This pipeline can serve as an implementation example for the scenario of proactive chatting with the elderly we discussed in this paper.

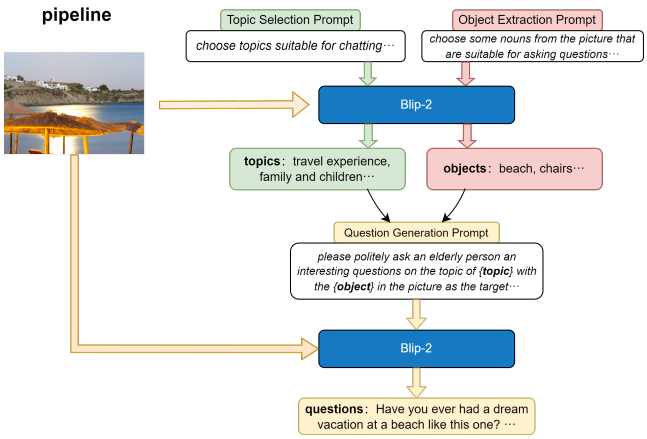

Fig. 4. The self-guided visual question generation pipeline.

V. EXPERIMENT AND EVALUATION

In this section, we will introduce the comparative experiments we designed and conducted to test the effectiveness of our proposed method. Simultaneously, our experiments can also be seen as an exploration of BLIP-2's capabilities in the task of question generation. However, it is worth noting that the method we propose is not exclusive to BLIP-2 and is applicable to any generative model capable of accepting both image and text inputs.

A. Partition of the Dataset

We randomly shuffle ChatVQG and divide the training set and test set in a ratio of 9:1.

B. Training Methods

We used two training methods to train the BLIP-2 model and compared their training effects.

(1) **Original:** In the paper of BLIP-2, the authors proposed a training method of freezing the large language model and training the visual encoder and QFormer for image-text generation. So we call such method as **original**.

(2) **Low-Rank Adaptation(LoRA):** It is an efficient large model fine-tuning strategy proposed by Microsoft [21]. It replaces the complete parameter update matrix with the product of two low-rank matrices.

C. Implementation

We use PyTorch and the transformers library developed by huggingface to train the model on two NVIDIA A6000s with 48G memory. Due to limited GPU resources, we need to adopt some memory optimization strategies such as half-precision and INT8 quantization [22] to ensure the execution of the training process.

D. BLIP-2 Variant Selection

As mentioned above, we have selected the BLIP-2 Flan T5 series model, which consists of two specific models:

- BLIP-2 ViT-G FlanT5$_{XL}$ with 4B parameters.
- BLIP-2 ViT-G FlanT5$_{XXL}$ with 12B parameters.

E. Automatic Evaluation Metrics

(1) **Accuracy:** We use accuracy to evaluate the model's ability to select topics. If the model generates the name of the topic during topic selection, it is considered that the model has selected the topic.

(2) **ROUGE** [23] is primarily employed in the context of language generation tasks to measure the similarity between the generated text and reference text.

(3) **Semantic Similarity:** As we mentioned before, the question generation task is an open-domain generation task. Therefore, we employ Sentence BERT [24] to calculate the semantic similarity between the model output and the reference.

(4) **Topic Relevance:** We use ChatGPT to automatically determine whether the questions generated by the model belong to the question topic specified by the input.

F. Evaluation of Topic Selection and Guided Question Generation

We use the automatic evaluation metrics introduced above to evaluate the performance of two selected BLIP-2 models trained with different training methods using our proposed method on ChatVQG for both topic selection task and topic-and-object-aware question generation. As shown in Table IV, BLIP-2 fine-tuned with LoRA outperforms the model trained with the original training method on all evaluation metrics.

TABLE IV
BASELINE EXPERIMENT RESULTS

Model	Accuracy	ROUGE-L	Similarity	Relevance
BLIP-2$^{\text{original}}_{\text{XL}}$	57.36	36.69	68.02	68.45
BLIP-2$^{\text{original}}_{\text{XXL}}$	56.02	36.50	67.00	59.63
BLIP-2$^{\text{lora}}_{\text{XL}}$	**61.19**	38.24	**69.96**	71.69
BLIP-2$^{\text{lora}}_{\text{XXL}}$	58.81	**38.50**	69.65	**74.01**

TABLE V
GUIDING LABELS EFFECTIVENESS EXPERIMENT RESULTS

Baseline	ROUGE-L	Similarity
BLIP-2$^{\text{original-unguided}}_{\text{XL}}$	30.91	52.05
BLIP-2$^{\text{original-guided}}_{\text{XL}}$	**36.69**	**68.02**
BLIP-2$^{\text{original-unguided}}_{\text{XXL}}$	30.54	52.35
BLIP-2$^{\text{lora-guided}}_{\text{XXL}}$	**36.50**	**67.00**
BLIP-2$^{\text{lora-unguided}}_{\text{XL}}$	30.59	51.59
BLIP-2$^{\text{lora-guided}}_{\text{XL}}$	**38.24**	**69.96**
BLIP-2$^{\text{lora-unguided}}_{\text{XXL}}$	31.14	51.20
BLIP-2$^{\text{lora-guided}}_{\text{XXL}}$	**38.50**	**69.65**

We conducted an experiment to emphasize the importance of guiding labels in ChatVQG. We compared two training methods for each model: one using only images and another using both images and guiding labels as inputs. Afterward, we assessed the quality of the generated content from the same model trained with these two approaches. As shown in Table V, each model trained with guiding labels clearly outperforms the one without guiding labels in every metric.

G. Human Evaluation of Self-Guided Pipeline

Finally, we employ human evaluation to assess the model's question generation quality within real-life application scenarios. We input the images from the test set into the self-guided pipeline and collect all the questions generated by the pipeline, along with question topics and question objects selected by the model. Evaluators will assign a score to each question based on the following evaluation criteria, with a maximum score of 5 points.

Suitability: Are the topic and object chosen by the model appropriate for the image?

Relativity: Is the question generated by the model highly relevant to its corresponding topic and object?

Politeness: Is the question generated by the model polite enough to ask to the elderly?

Engagingness: Is the question generated by the model engaging enough to be answered?

Subsequently, we calculate the average scores for all test images across each criterion, as illustrated in Fig. 5.

H. Result Analysis

Based on the experimental results, we can draw the following conclusions:

(1) **LoRA is better than freezing layers when GPU resources are limited.** As shown in Table IV, the model trained using LoRA is stronger than the model trained using the original training method, which freezes all layers in LLM,

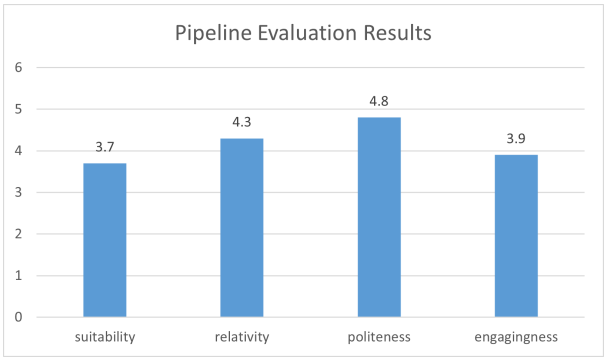

Fig. 5. Human evaluation results of the pipeline.

in all metrics. This is mainly because LoRA efficiently fine-tunes a small number of parameters per layer, aiding effective learning from the training data. In contrast, the original method demands more training data and additional computing resources. To train, we need to employ mixed or half-precision, which could slightly affect the training results. This finding can be valuable for BLIP-2 users facing GPU resource constraints.

(2) **The guiding labels we proposed play a significant role in improving model controllability, which, in turn, is very helpful for enhancing the quality of question generation and ensuring accurate evaluation.** As shown in Table V, the obvious difference between the two settings is self-evident. However, we believe that the lower scores obtained by unguided models do not mean that their generative capabilities are weaker because we are comparing models with exactly the same structure and number of parameters. This shows that when evaluating open-domain generation tasks, not using guide labels to control its generation direction does not objectively reflect the generation quality of the model.

(3) **The self-guided pipeline we proposed essentially fulfills the requirement of proactively asking questions to the elderly to initiate a conversation.** As shown in Fig. 5, the average score on each criterion in the human evaluation has surpassed 3 points, with topic relevance and politeness both exceeding 4 points. This demonstrates that the questions generated by the pipeline are topical and appropriate for asking the elderly. The good engagingness indicates that the questions raised can facilitate multi-round conversations.

REFERENCES

[1] S. Valtolina and L. Hu, "Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness," in *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, 2021, pp. 1–5.

[2] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, "Generating natural questions about an image," *arXiv preprint arXiv:1603.06059*, 2016.

[3] M.-H. Yeh, V. Chen, L.-W. Ku *et al.*, "Multi-vqg: Generating engaging questions for multiple images," *arXiv preprint arXiv:2211.07441*, 2022.

[4] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*. Springer, 2014, pp. 529–545.

[5] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[6] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[7] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2596–2604.

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

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[10] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4894–4902.

[11] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[12] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[13] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[14] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions," *arXiv preprint arXiv:2303.06594*, 2023.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[17] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1931–1942.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[19] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[22] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm. int8 (): 8-bit matrix multiplication for transformers at scale," *arXiv preprint arXiv:2208.07339*, 2022.

[23] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.