Explicitly Guided Difficulty-Controllable Visual Question Generation

Jiayuan Xie[1*], Mengqiu Cheng[6*], Xinting Zhang[4], Yi Cai[2], Guimin Hu[3†], Mengying Xie[5], Qing Li[1]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[2]School of Software Engineering, South China University of Technology, Guangzhou, China
[3] Department of Computer Science, University of Copenhagen, Denmark
[4] Department of Mathematics, The University of Hong Kong, Hong Kong SAR, China
[5] College of Computer Science, Chongqing University, China
[6] Guangdong Neusoft University, Foshan, China
jiayuan.xie@polyu.edu.hk

Abstract

Visual question generation (VQG) aims to generate questions from images automatically. While existing studies primarily focus on the quality of generated questions, such as fluency and relevance, the difficulty of the questions is also a crucial factor in assessing their quality. Question difficulty directly impacts the effectiveness of VQG systems in applications like education and human-computer interaction, where appropriately challenging questions can stimulate learning interest and improve interaction experiences. However, accurately defining and controlling question difficulty is a challenging task due to its multidimensional and subjective nature. In this paper, we propose a new definition of the difficulty of questions, i.e., being positively correlated with the number of reasoning steps required to answer a question. For our definition, we construct a corresponding dataset and propose a benchmark as a foundation for future research. Our benchmark is designed to progressively increase the reasoning steps involved in generating questions. Specifically, we first extract the relationships among objects in the image to form a reasoning chain, then gradually increase the difficulty by rewriting the generated question to include more reasoning sub-chains. Experimental results on our constructed dataset show that our benchmark significantly outperforms existing baselines in controlling the reasoning chains of generated questions, producing questions with varying difficulty levels.

Introduction

Visual Question Generation (VQG) aims to automatically generate questions from images, a task that has garnered significant attention in the vision and language communities in recent years (Xie et al. 2023). The ability to generate questions automatically has substantial implications across various domains, such as providing dynamic demonstrations in children's education (Kunichika et al. 2004) and initiating conversations in chatbots (Xie et al. 2024b). However, existing studies primarily focus on the quality of generated questions, such as fluency and relevance, while neglecting the crucial factor of question difficulty. Educational research (Ha et al. 2019) indicates that controlling question difficulty is essential for effective learning, as appropriately challenging questions can better assess students' comprehension levels and provide personalized learning experiences. Questions with suitable difficulty can stimulate learning interest, enhance cognitive training outcomes, and improve user experience in intelligent human-computer interaction. Therefore, it is crucial to focus on generating questions with appropriate levels of difficulty to maximize the effectiveness and applicability of VQG systems.

Although the mainstream VQG task (Xie et al. 2021) has made significant research progress, the challenge of controlling question difficulty remains substantial. This is primarily because formally defining question difficulty is inherently subjective and involves multiple complexities (Bramley 2011). To the best of our knowledge, the only existing work on difficulty-controllable VQG (DVQG) is by Chen et al. (2023). Their approach defines question difficulty based on whether certain existing visual question answering (VQA) models can correctly answer the questions. In their definition, a question is considered "easy" if VQA models can answer it correctly and "hard" if they cannot. However, this work has three limitations. Firstly, it only provides two levels of difficulty (easy and hard), failing to capture the nuanced spectrum of question difficulty. Secondly, this definition is heavily dependent on the chosen VQA models, thus lacking generalizability. More importantly, it lacks interpretability regarding what makes a question difficult and how the difficulty changes, which is crucial for practical applications in education and cognitive training. Thus, there is an urgent need for a more reasonable and robust definition of difficulty for the DVQG task.

To address the aforementioned three limitations, we propose a new difficulty level definition for the DVQG task. Inspired by research about multi-hop questions (Fang et al. 2020; Sui et al. 2022), we define difficulty levels based on the number of reasoning steps required to answer the questions. As the number of reasoning steps increases, the questions become more difficult, thus overcoming the first limitation of having only "easy" or "hard" levels. As shown in the example in Figure 1, answering $\mathbf{Q}_1$ only requires a one-hop chain of reasoning, i.e. $N_0 \rightarrow N_1$; While answering $\mathbf{Q}_2$, the two hops of the reasoning chain $N_0 \rightarrow N_1$ and $N_1 \rightarrow N_2$ need to be considered step-by-step, which increases the complex-

[*]equal contribution
[†]Corresponding author

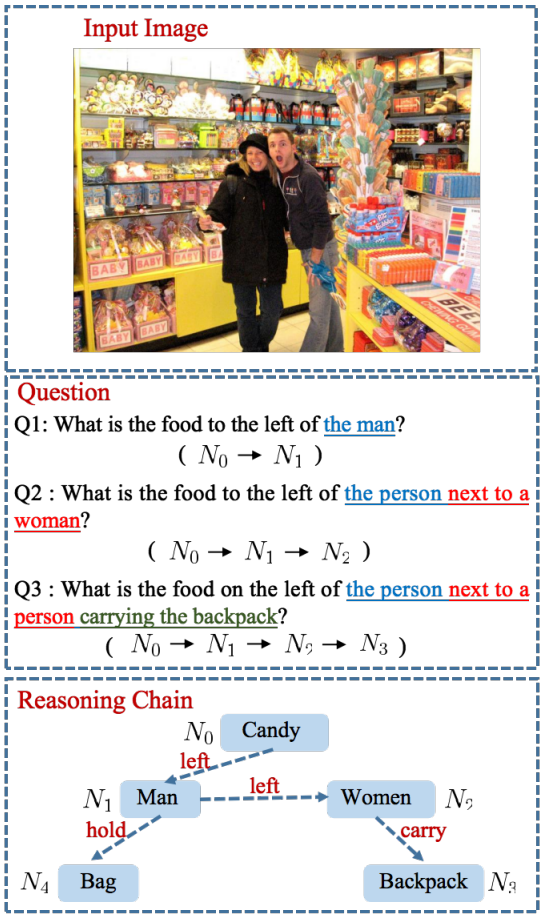

Figure 1: Visual question samples from the GQA dataset. $\mathbf{Q}_1$ corresponds to $N_0 \rightarrow N_1$; $\mathbf{Q}_2$ corresponds to $N_0 \rightarrow N_1 \rightarrow N_2$; $\mathbf{Q}_3$ corresponds to $N_0 \rightarrow N_1 \rightarrow N_2 \rightarrow N_3$.

ity of the question. According to our experiments, existing VQA systems (Norcliffe-Brown, Vafeias, and Parisot 2018; Anderson et al. 2018; Xie et al. 2024a) perform significantly worse on multi-hop questions (e.g., $\mathbf{Q}_2$, $\mathbf{Q}_3$ in Figure 1) compared to 1-hop ones (e.g., $\mathbf{Q}_1$ in Figure 1), which increases the reliability of using the number of reasoning steps to define the difficulty. Based on this new definition, we address the limitation of dependency on specific VQA models by grounding our difficulty definition in reasoning complexity rather than model performance. Moreover, this step-by-step reasoning process makes it clear what factors contribute to the increased difficulty, thus addressing the third limitation regarding the interpretability of question difficulty.

To this end, we propose a multi-step question generation model (MultiStepGen), which explicitly controls the difficulty of the generated questions based on the reasoning chain information in the given image. Our model contains three components, i.e., reasoning chain extractor, visual feature extractor, and controllable rewriting module. Compared with existing methods on VQG that merely utilize visual features, we introduce the reasoning chain information through the reasoning chain extractor to provide more instructive for question generation. When a reasoning chain contains rich information (i.e., contains multiple sub-chains), we observe that existing generation models ignore some sub-chains and fail to generate logically rigorous multi-hop questions. Therefore, the controllable rewriting module adopts a mechanism to rewrite a question involving N sub-chains into a more complex question involving N+1 sub-chains. We can generate multi-hop questions by progressively rewriting a question, and ensure that each sub-chain information can be utilized. As shown in Figure 1, we first utilize the visual features and the selected reasoning chain $N_0 \rightarrow N_1$ related to the given answer "Candy" to generate an initial 1-hop question $\mathbf{Q}_1$. We then rewrite $\mathbf{Q}_1$ to 2-hop question $\mathbf{Q}_2$ by a longer reasoning chain (i.e. $N_0 \rightarrow N_1 \rightarrow N_2$), which contains more complex reasoning chain than $\mathbf{Q}_1$. Similarly, we can further increase the generated question difficulty level by step-by-step increasing the sub-chains, e.g., the 3-hop question $\mathbf{Q}_3$.

In summary, our contributions are as follows:

- To address the limitations of existing research on DVQG, our work defines question difficulty as the number of reasoning steps required to answer the question, breaking away from the traditional binary classification of easy and hard, and enabling explicit control over the difficulty of generated questions.
- Unlike existing VQG models that merely extract visual information, we introduce the incorporation of reasoning chain information, which provides the foundation for controlling the reasoning steps involved in question generation. Additionally, we designed a rewriting mechanism that dynamically controls question difficulty by progressively increasing the sub-chains of the given reasoning chain. This step-by-step rewriting can generate multi-hop questions and ensures that each sub-chain's information is fully utilized.
- According to our proposed difficulty definition, we construct our DVQA dataset from the GQA dataset (Hudson and Manning 2019) to evaluate model performance. Experimental results show that our proposed framework outperforms existing state-of-the-art models in both automatic and human evaluations, and can controllably generate questions with the required number of reasoning steps.

Related Work

Most studies tackle the VQG task with deep neural networks. Mostafazadeh et al. (2016) build three VQG datasets and propose an end-to-end neural model to tackle the task of VQG. Considering that previous research mainly generates generic and uninformative questions, Krishna et al. (2019) argue that a good question should aim to expect a specific target and propose a model that maximizes the mutual information between the generated question, the image, and the target answer. Xu et al. (2021) propose Radial-GCN model based on the object-level features of an image, which captures the relations between the answer area and the most relevant image region. Since a target answer may be related to more than one image region, Xie et al. (2021) first extract one or more image regions related to the answer based on image object-level features, and then simultaneously focus on multiple image regions for question generation. Xie et

al. (2022) propose combining additional knowledge to generate questions that beyond visual features, which can enrich the content involved in generating questions. Their studies focus on generating questions solely based on images without considering the issue of difficulty. Fang et al. (2024) propose to use an expert mechanism to extract multiple key different objects in an image and then generate diversity questions with different key objects. The difficulty is an important factor in measuring the quality of generated questions. To the best of our knowledge, Chen et al. (2023) is the first to propose DVQG, and they argue that difficulty can be used as an indicator to guide question generation. Inspired by this, this work explores another dimension of difficulty, which is the difficulty metric as the length of the reasoning chain required to answer the question.

Dataset Construction

Existing datasets on VQG (Anderson et al. 2018; Goyal et al. 2017) are insufficient to support the evaluation of this task, primarily due to the new definition of question difficulty. Specifically, we define question difficulty as the number of reasoning steps required to answer the question. In detail, these datasets cannot contain multiple questions with consecutive reasoning steps, and thus fail to train an effective model or verify whether the model can effectively generate questions that meet the proposed difficulty definition.

As we know, constructing a dataset suitable for our definition of the DVQG task from scratch is labor-intensive. Thus, we propose to perform secondary processing on the existing GQA dataset (Hudson and Manning 2019) that has been used for traditional VQG. The GQA dataset automatically constructs diverse questions involving various reasoning skills mainly through the visual genome scene graph structure (Johnson et al. 2015; Krishna et al. 2017), which is a dataset for real-world visual reasoning questions answering. Each sample of the GQA dataset mainly consists of an image, an answer, and a list of questions related to the image and the answer. In addition, a series of reasoning steps required to answer each question is included. We process the GQA dataset to suit our task in the following two steps, i.e., preprocessing and question pair construction.

Preprocessing The reasoning steps in the GQA dataset mainly include the following situations, i.e., "select", "filter", and "relate". We choose to retain samples containing the "relate" type and filter out boolean questions. The reason for selecting "relate" type samples is that they typically involve understanding and reasoning about the relationships between multiple objects within an image, which aligns closely with our goal of constructing multi-step reasoning chains. By focusing on "relate" samples, we ensure that the questions in our dataset require more complex reasoning processes, which are essential for validating these difficulty-controlled question generation approaches.

On the other hand, we filter out boolean questions (i.e., "yes/no" questions) because we need to determine the required reasoning chain based on the answer to the question, and boolean questions fail to provide this information.

Question Pair Construction Given the need to generate questions with different reasoning steps, each sample in our dataset must contain a set of questions with continuous reasoning chains for training and validation. However, the original GQA dataset often fails to meet this requirement. Therefore, we needed to reasonably construct and pair questions based on the preprocessed data to form valid question pairs.

In detail, we retain merely 1-hop and 2-hop question pairs. The reason for this choice is that 1-hop and 2-hop reasoning chains strike a reasonable balance between complexity and controllability. These question pairs are sufficient to test the model's performance at different reasoning difficulties while ensuring the continuity and coherence of the constructed questions. Moreover, multi-hop questions (more than 2 hops) significantly increase complexity, which may introduce too much noise and uncertainty, potentially affecting the model's training and evaluation results. Thus, we prioritized 1-hop and 2-hop question pairs as the foundation for our dataset construction. In cases where a 2-hop question cannot find a directly corresponding continuous 1-hop question in the dataset, we construct a set of questions with continuous reasoning chains. Specifically, we use the included 1-hop reasoning chain to generate the corresponding 1-hop questions through the ChatGPT (OpenAI 2023).

Additionally, we perform manual checks on the generated questions to ensure data quality and the coherence of reasoning chains. The manual checking process includes the following steps, i.e., i) Reasoning Chain Consistency Verification: Check whether the generated question sets adhere to the expected reasoning chain structure, ensuring logical continuity from 1-hop to 2-hop questions. ii) Semantic Accuracy Check: Perform a semantic analysis of the generated 1-hop questions to ensure they are related to the corresponding 2-hop questions.

Model

In this DVQG task, given an image I, a target answer A and a specific difficulty level d, our goal is to generate a question $\mathbf{Q}_d$ related to the image I and its answer A, where $\mathbf{Q}_d$ requires d reasoning steps to answer. The overall framework of our multi-step question generation model (MultiStepGen) can be seen in Figure 2, which consists of three components, i.e., reasoning chain selection (RCS), visual feature extractor (VFE), and controllable rewriting module (CRM). First, the RCS constructs a scene graph S corresponding to a given image, and selects a relationship chain T_d: $N_0 \rightarrow N_1 \rightarrow ... N_d$ ($T_d \in S$) related to the given answer A as the reasoning chain of generated question. Then, with the reasoning chain T_0: $N_0 \rightarrow N_1$ and the image information extracted from the VFE as input, the CRM produces an initial simple question $\mathbf{Q}_1$. Finally, the next step of the CRM iteratively generates more complex question $\mathbf{Q}_i$ ($i = 2, 3, ..., d$) based on the $\mathbf{Q}_{i-1}$ of the previous step and a relationship chain T_{i-1}: $N_0 \rightarrow N_1 ... \rightarrow N_i$ ($T_{i-1} \in S$).

Reasoning Chain Selection

To extract the appropriate reasoning chain from the scene graph of an image, we mainly include two steps, i.e., (a) Scene Graph Construction and (b) Answer-aware Selector.

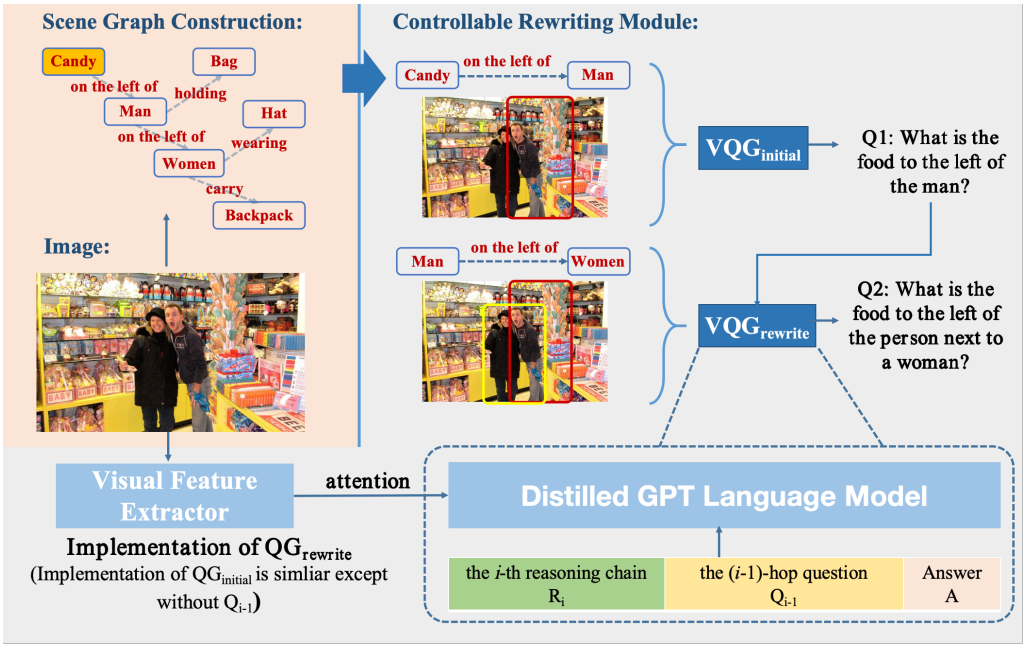

Figure 2: Overview of our model MultiStepGen. The model is to generate the multi-hop questions step-by-step.

Scene Graph Construction Following Krishna et al. (2017), we annotate each image with a dense scene graph S, which contains the objects in the graph and their attributes and relationships. Each node in the graph represents an object, and two nodes with a relationship can be connected by a directed edge, where the edge describes the relationship between them. As shown in Figure 1, an object "Candy" (N_0) and an object "Man" (N_1) are connected by the relationship "on the left of", i.e., $N_0 \rightarrow N_1$. We utilize coreference resolution (Lee et al. 2017) to merge different relations of the same object, e.g., $N_0 \rightarrow N_1$ and $N_1 \rightarrow N_2$ to $N_0 \rightarrow N_1 \rightarrow N_2$.

Answer-aware Selector We select a relationship chain T_d consisting of (d+1) nodes from the scene graph S to generate questions, where the head node N_0 in the T_d is required to be related to the given answer. Therefore, we need to extract an object most relevant to the given answer in the image as node N_0. Specifically, we first utilize the BERT (Devlin et al. 2018) representation to obtain the features of the answer and object labels. Then, we use cosine similarity to compute pairwise semantic relevance scores between each object and its corresponding answer. Finally, we select the object with the highest correlation score as N_0 and its related chain T_d as our reasoning chain.

Visual Feature Extractor

We employ the pre-trained CLIP ViT-B/16 (Radford et al. 2021) as the visual feature extractor. On one hand, CLIP is a model that pre-trained with massive image-text pairs, which can ensure the extracted visual features are semantically aligned with the text content. On the other hand, CLIP has shown powerful capabilities for capturing rich visual semantics. Based on the two aspects, we adopt CLIP to extract visual features.

In our process, the image is first resized to a standard resolution of 224 $\times$ 224 pixels. This resizing ensures uniformity across all input images, facilitating consistent feature extraction. After resizing, the image is divided into P = 14 $\times$ 14 = 196 patches, each with a size of 16 $\times$ 16 pixels. This patching process breaks down the image into manageable segments (Cheng and Sun 2024), allowing the model to focus on finer details within each region. Each of these 196 patches is then passed through the CLIP visual encoder, which computes a visual feature vector v_p for each patch. These feature vectors are highly representative of the visual content within each patch and are essential for downstream tasks that require precise image understanding.

As a result, the collection of visual features from all patches can be denoted as $V = \{v_p\}_{p=1}^{196}$. This set of features provides a comprehensive representation of the image, capturing the various elements and their interactions within the visual scene.

Controllable Rewriting Module

Initial Question Generation Considering the powerful generative capabilities of large-scale language models (LLMs) (Xie et al. 2024c; Shen and Tang 2024), we use it as the question generation model. Specifically, the initial step employs a fine-tuned GPT-2 (Radford et al. 2019; Liu et al. 2024) as the generation model, which has been pre-trained on a large-scale image captioning dataset to ensure its effectiveness in visual and language tasks. We feed the answer and the first-hop reasoning chain into the decoder of GPT-2 to generate a 1-hop question.

We formalize the input sequence by merging the first-hop reasoning chain as the template: "This is a 1-hop question, the reasoning chain is R_1, and the answer is A". The ques-

tion $\mathbf{Q}_1$ is then generated in an autoregressive manner, beginning with the start-of-sequence token BOS, followed by the content of the 1-hop question, and ending with the end-of-sequence token EOS.

To ensure that the generated question aligns with the fused visual features, we use the hidden state h_i from the GPT-2 at each time step as the query, and the visual features v_i as the keys and values, applying the vanilla attention mechanism (Vaswani et al. 2017) for fusing text and visual features.

The model is trained using a cross-entropy objective to generate a sequence of T words, $y = \{y_1, y_2, ..., y_T\}$, as the question. The goal of the training process is to minimize the negative log-likelihood, thereby improving the accuracy and relevance of the generated questions. The formula is calculated as follows:

$$L = -\sum_{\theta=1}^{T} \log p(y_\theta | y_{<\theta}), \tag{1}$$

where $y_{<\theta}$ denotes the words before the θ-th word.

Complex Question Generation After the initial decoder generates a 1-hop simple question $\mathbf{Q}_1$, the rewritten decoder aims to generate a more complex multi-hop question. Different from the generation of the initial question, we adjust the input of GPT-2. Specifically, we introduce the $\mathbf{Q}_{(N-1)}$ from the previous step as the input, i.e., "This is a N-hop question, the reasoning chain is R_N, the (N-1)-hop question is $\mathbf{Q}_{(N-1)}$ and the answer is A".

Experiment Settings

Dataset

The dataset we constructed in this paper contains 220,657 question pairs. Specifically, 80% of our dataset is used as a training set, 10% as a validation set, and 10% as a test set.

Baseline Methods

To evaluate the effectiveness of our framework, we compare our models with several baselines. Our experiments mainly consider two types of models: existing baseline methods and the variants of our methods. The baselines are as follows:

• **GRNN** (Mostafazadeh et al. 2016) is a baseline model for visual question generation. It uses VGGNet as the image encoder and GRU as a decoder to generate questions solely based on images while neglecting the answer information.

• **IM-VQG** (Krishna, Bernstein, and Fei-Fei 2019) utilizes a ResNet model to encode an image, and combines the information of the answer and the answer category to generate a question.

• **Radial-GCN** (Xu et al. 2021) extracts object-level features in an image, and utilizes the radial GCN to focus on an object most relevant to the answer for question generation.

• **MOAG** (Xie et al. 2021) simultaneously focuses on one or more objects in the image that are relevant to the answer for question generation.

• **MS-VQG** (Fang et al. 2024) focuses on the objects in the reasoning chain for question generation.

• **ChatGPT** (OpenAI 2023) directly generates questions based on answers and reasoning chains.

The variants contain i) **MultiStepGen w/o VFE**, which ignores image information for question generation; ii) **MultiStepGen w/o CRM**, which ignores the step-by-step generation process and trains and predicts all data together; iii) **MultiStepGen w/o A**, which ignores the answer information for question generation.

Evaluation

Automatic Metrics To compare our proposed model with baseline models, we report commonly-used metrics in text generation, i.e., BLEU (1 to 4) (Papineni et al. 2002), ROUGE$_L$ (Lin 2004), METEOR (Denkowski and Lavie 2014) and CIDEr (Vedantam, Zitnick, and Parikh 2015).

Human Evaluation Criteria In addition to the automatic evaluation, we invite five volunteers with a rich educational experience to judge the quality of questions generated by different models based on 200 samples (Xing et al. 2017; Fan et al. 2018). Volunteers refer to the following criteria to judge the quality of the generated questions: Fluency (F) measures the grammatical correctness and fluency of the generated question; Relevance (R) assessment whether the generated question is relevant to the image and the target answer; Difficulty (D) mainly reflects the difficulty level of the generated question, whether a longer chain of reasoning is required; Answerability (A) measures whether a question can be answered by the given answer. where F, R, and D take values from $\{0, 1, 2\}$ (higher values indicate better-generated questions), while A takes values from $\{0$ or $1\}$ (1 means answerable, 0 means not answerable).

Experimental Details

Our model is implemented using the PyTorch framework and trained on a single GTX2080 Ti GPU. For the visual encoder, we employ a CLIP model built with a ViT-B/16 Transformer architecture, which has been pre-trained on publicly accessible image-caption datasets (Radford et al. 2021). The GPT-2 model we use has been distilled and pre-trained on a large-scale collection of image-caption pairs (Sammani, Mukherjee, and Deligiannis 2022). The model is trained for up to 5 epochs utilizing the Adamax optimizer (Kingma and Ba 2015), with a batch size of 128 and a learning rate of 2×10^{-5}.

Results and Analysis

Comparison with Existing Models The first part of Table 1 shows the automatic evaluation results of our model MultiStepGen and baselines. We have several findings:

• In the experiments, all comparison models significantly outperformed GRNN, particularly with an increase of at least 8.99 in the BLEU 4 metric of 1-hop questions. This result indicates that incorporating constrained answers and reasoning chains provides crucial information that enhances the model's ability to focus on key details within the image, thereby generating more relevant questions.

• Aside from IMVQG and GRNN, the other models utilized object-level features for question generation, which resulted in noticeable performance improvements. This suggests that object-level features are instrumental in better cap-

	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	CIDER	METEOR	ROUGE
1-hop	GRNN-KB	25.59	16.39	9.27	5.29	0.46	10.17	26.08
	IMVQG	36.53	24.38	17.87	11.08	1.24	13.98	34.29
	VGQ-GCN	45.12	31.69	23.61	14.28	1.44	20.46	40.28
	VISUAL-BERT	56.37	43.35	25.87	18.38	1.57	23.82	56.59
	MS-VQG	48.33	34.23	25.51	17.72	1.55	22.62	48.18
	ChatGPT	62.54	49.45	38.94	31.66	3.02	36.27	58.29
	MultiStepGen w/o VFE	58.58	43.88	32.19	25.79	2.41	33.94	55.63
	MultiStepGen w/o CRM	60.27	44.74	34.65	26.13	2.29	34.49	55.39
	MultiStepGen w/o A	61.79	48.47	37.89	30.61	2.82	35.54	55.93
	MultiStepGen	**65.36**	**51.73**	**41.01**	**32.26**	**3.51**	**37.25**	**60.02**
2-hop	GRNN-KB	20.68	14.18	7.91	4.87	0.42	8.53	19.64
	IMVQG	25.59	18.13	14.87	10.44	0.92	14.54	20.23
	VGQ-GCN	30.52	29.32	20.36	11.78	1.14	16.46	24.28
	VISUAL-BERT	36.47	31.12	24.26	14.36	1.22	22.31	36.99
	MS-VQG	32.26	28.98	20.27	10.74	1.05	17.02	26.73
	ChatGPT	26.96	20.28	14.45	7.23	1.02	15.43	22.58
	MultiStepGen w/o VFE	39.54	30.24	21.39	15.18	1.18	20.46	44.25
	MultiStepGen w/o CRM	38.21	29.87	21.31	14.79	1.18	20.22	43.57
	MultiStepGen w/o A	45.86	32.28	22.25	16.23	1.24	20.87	45.58
	MultiStepGen	**46.68**	**34.39**	**24.76**	**18.01**	**2.11**	**23.27**	**47.54**

Table 1: Main automatic metrics results of baselines and our model on our DVQA dataset.

turing regions of the image that are relevant to the reasoning chain. Consequently, models leveraging these features were able to generate questions that were more accurately aligned with the reasoning chain, thus improving both the accuracy and logical consistency of the generated questions.

• When generating 1-hop questions, most baselines performed well. Among them, the ChatGPT model is particularly prominent, as the construction of some 1-hop questions utilizes the ChatGPT model. However, their effectiveness notably decreased when generating 2-hop questions, with ChatGPT experiencing the most significant drop in performance. This decline can be attributed to the relative simplicity of 1-hop questions, whereas 2-hop questions involve longer reasoning chains and more complex associations with the image content. In contrast, our approach not only leverages the powerful generation capabilities of LLMs but also integrates image information throughout the generation process. More importantly, our method enables the step-wise generation of questions with increasing complexity, ensuring the continuity of reasoning chains between questions, thereby enhancing both the accuracy and logical coherence in multi-hop question generation.

Ablation Study The second part of Table 1 shows the performance of our variant models. We find that:

• When the VFE is removed, the model's performance significantly declines, with the BLEU-4 dropping by nearly 15% in 2-hop questions. This indicates that visual information is crucial for maintaining the integrity of the question.

• Similarly, when the CRM is removed, the model's performance also saw a significant drop, with the BLEU 4 again decreasing by nearly 18% in 2-hop questions. This result suggests that the stepwise approach to question generation effectively ensures the controllability of the question content (Zhang, Wang, and Zhang 2024), as each step builds upon the previous one, incorporating minimal new content while maintaining consistency.

• When the answer is removed, the model's performance experiences a slight decrease, with the BLEU 4 dropping by nearly 10%, though this change is not as pronounced. This suggests that the content provided by the answer overlaps with the reasoning chain, resulting in a less significant impact on overall performance.

	F	D	R	A
VisualBERT (1-hop)	1.80	1.41	1.36	0.31
VisualBERT (2-hop)	1.62	1.44	1.15	0.27
MultiStepGen (1-hop)	**1.97**	1.02	**1.79**	**0.80**
MultiStepGen (2-hop)	**1.97**	**1.98**	1.70	0.71

Table 2: The human evaluation results.

Human Evaluation Table 2 shows the results of human evaluation, where we select VisualBERT, the best-performing baselines, as the benchmark for comparison. We calculated the standard deviation of the evaluators' scores on the human evaluation criteria, with all values being below 0.15. This indicates that the human evaluation results are highly reliable. In terms of "Fluency", both VisualBERT and MultiStepGen receive high scores, reflecting the strong language modeling capabilities of neural models.

For the "Difficulty", we observed that the difficulty gap

Model	Up-Down	DFAF	Counter	Graph
DGN (easy)	40.33	39.67	43.32	40.83
DGN (hard)	40.10	39.78	42.16	40.13
Ours (1-hop)	42.76	42.59	47.36	44.96
Ours (2-hop)	39.34	39.74	43.23	40.22
GQA (1-hop)	41.47	40.94	46.21	43.73
GQA (2-hop)	39.74	40.07	44.04	40.97

Table 3: Different VQA models evaluate the generated questions, which are the results generated by DGN, our model MultiStepGen with difficulty control.

between the 1-hop and 2-hop questions generated by MultiStepGen was greater than that of VisualBERT, indicating that the proposed MultiStepGen model is more effective at controlling the difficulty of the generated questions.

Moreover, MultiStepGen outperformed VisualBERT in both "Relevance" and "Answerability", demonstrating that our approach can better leverage reasoning chains and answer information to generate higher-quality questions.

Difficulty control experiment Inspired by the research of Chen et al. (2023), we believe that the difficulty of the question can be reflected in the answering results of the respondents. Therefore, we use VQA models to simulate these respondents to evaluate the difficulty of the question, where the evaluation metric is the accuracy rate. Specifically, we select four different existing VQA models for evaluation, i.e., Up-Down (Anderson et al. 2018), DFAF (Gao et al. 2019), Counter (Zhang, Hare, and Prügel-Bennett 2018), Graph (Norcliffe-Brown, Vafeias, and Parisot 2018).

For a fair comparison, we also introduce the DGN model (Chen et al. 2023) to generate hard and easy questions. The results of the difficulty control experiment are shown in Table 3, and we find that: i) We observe that the VQA models performed better on 1-hop questions (e.g., GQA (1-hop) and MultiStepGen (1-hop)) compared to 2-hop questions (e.g., GQA (2-hop) and MultiStepGen (2-hop)). This finding suggests that our model can effectively generate questions with varying difficulty levels based on the reasoning steps involved; ii) GQA (1-hop) and GQA (2-hop) are manually constructed datasets where the accuracy of VQA models is not influenced by the quality of the questions. This further indicates that, for VQA models, 2-hop questions are more challenging to answer than 1-hop questions.

In the case of the DGN model, the performance difference between easy and difficult generated questions in this VQA task is not significant. For instance, in the Counter model, the accuracy of easy questions was only +0.16 higher than that of difficult questions. In contrast, our model shows greater improvement because it is better suited for generating questions with clear differences in difficulty.

Case Study

Figure 3 illustrates the visual questions generated by our MultiStepGen and VisualBERT. We find that: i) The questions generated by VisualBERT show instances of repeated

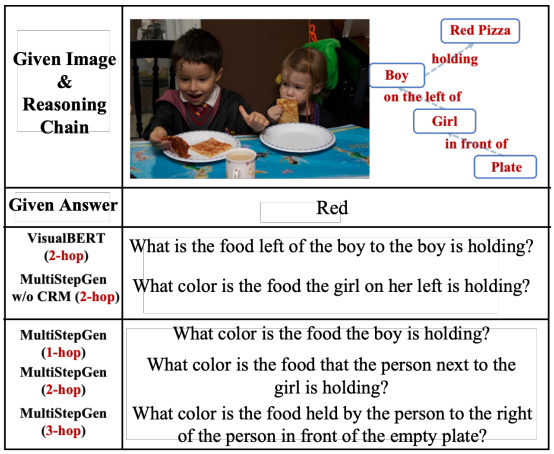

Figure 3: Case study of sample output questions.

or incorrect triplets, indicating that the model struggles to accurately capture relationships between objects. In contrast, our model leverages the powerful capabilities of GPT to more accurately capture the relationships of these objects; ii) We also observed that the questions generated by MultiStepGen (1-hop) contain only one relationship triplet, making them less challenging; whereas the questions generated by MultiStepGen (2-hop) include two relationship triplets, requiring multiple reasoning steps, thus being more difficult. This demonstrates that our MultiStepGen can effectively control the difficulty of question generation based on the number of reasoning steps. iii) MultiStepGen w/o CRM overlooks the parts of the subchain content in the generation of 2-hop questions. This further supports the effectiveness of the step-wise generation approach in helping the model better capture the content of reasoning chains; iv) Based on training data containing 1-hop and 2-hop questions, the proposed model is able to generate some high-quality 3-hop questions, demonstrating the scalability of our model. Besides, we extracted 100 multi-hop ($\geq$ 3-hops) questions from the original GQA dataset for validation, and the BLEU 4 score is 17.23. This indicates that our model can ensure the generated effect while increasing the number of chains.

Conclusion

In this paper, we define the difficulty of visual questions as the number of reasoning steps required to answer them. Based on this definition, we propose a new dataset and an iterative question generation model with controllable reasoning steps, laying the foundation for future research. Using the constructed dataset, we develop a model called MultiStepGen that learns how to rewrite 1-hop questions into 2-hop questions, where the number of hops represents the number of subchains in the question. Our model not only performs well in generating 2-hop questions but also effectively scales to more complex multi-hop questions. The experimental results demonstrate that the proposed model outperforms strong baseline models in both key metrics and human evaluations, and it can flexibly generate questions of varying difficulty based on the reasoning steps.

Acknowledgements

This research is supported by the Science and Technology Planning Project of Guangdong Province (2020B0101100002), the National Natural Science Foundation of China (62076100, 62476097), the Fundamental Research Funds for the Central Universities, South China University of Technology (x2rjD2240100), Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), the China Computer Federation (CCF)-Zhipu AI Large Model Fund, a grant from HK RGC Theme-based Research Scheme (PolyU No.: T43-513/23-N), the Postdoctoral Fellowship Program of CPSF under Grant GZB20230912, in part by the Chongqing Postdoctoral Innovation Talents Support Program under Grant CQBX202324.

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.

Bramley, T. 2011. Subject difficulty-the analogy with question difficulty.

Chen, F.; Xie, J.; Cai, Y.; Lin, Z.; Li, Q.; and Wang, T. 2023. Graph convolutional network for difficulty-controllable visual question generation. *World Wide Web*, 26(6): 3735–3757.

Cheng, S.; and Sun, H. 2024. SPT: Sequence Prompt Transformer for Interactive Image Segmentation. arXiv:2412.10224.

Denkowski, M. J.; and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *ACL workshop*, 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*.

Fan, Z.; Wei, Z.; Wang, S.; Liu, Y.; and Huang, X. 2018. A Reinforcement Learning Framework for Natural Question Generation using Bi-discriminators. In *COLING*, 1763–1774.

Fang, W.; Xie, J.; Liu, H.; Chen, J.; and Cai, Y. 2024. Diverse Visual Question Generation Based on Multiple Objects Selection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(6): 1–22.

Fang, Y.; Sun, S.; Gan, Z.; Pillai, R.; Wang, S.; and Liu, J. 2020. Hierarchical Graph Network for Multi-hop Question Answering. In *EMNLP*, 8823–8838.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C. H.; Wang, X.; and Li, H. 2019. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *CVPR*, 6639–6648.

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*, 6325–6334.

Ha, L. A.; Yaneva, V.; Baldwin, P.; and Mee, J. 2019. Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam. In *BEA@NAACL-HLT*, 11–20.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 6700–6709.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *CVPR*, 3668–3678.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2019. Information Maximizing Visual Question Generation. In *CVPR*, 2008–2018.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.*, 123(1): 32–73.

Kunichika, H.; Katayama, T.; Hirashima, T.; and Takeuchi, A. 2004. Automated question generation methods for intelligent English learning systems and its evaluation. In *Proc. of ICCE*.

Lee, K.; He, L.; Lewis, M.; and Zettlemoyer, L. 2017. End-to-end Neural Coreference Resolution. In *EMNLP*, 188–197.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL workshop*, 74–81.

Liu, H.; Wang, G.; Xie, J.; Chen, J.; Fang, W.; and Cai, Y. 2024. Knowledge-Guided Cross-Topic Visual Question Generation. In *Proc. of LREC-COLING 2024*, 9854–9864.

Mostafazadeh, N.; Misra, I.; Devlin, J.; Mitchell, M.; He, X.; and Vanderwende, L. 2016. Generating Natural Questions About an Image. In *ACL*.

Norcliffe-Brown, W.; Vafeias, S.; and Parisot, S. 2018. Learning Conditioned Graph Structures for Interpretable Visual Question Answering. In *NeurIPS*, 8344–8353.

OpenAI. 2023. ChatGPT (September 25 Version) [Large language model].

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sammani, F.; Mukherjee, T.; and Deligiannis, N. 2022. NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks. In *CVPR*, 8322–8332.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sui, Y.; Feng, S.; Zhang, H.; Cao, J.; Hu, L.; and Zhu, N. 2022. Causality-aware Enhanced Model for Multi-hop Question Answering over Knowledge Graphs. *Knowl. Based Syst.*, 250: 108943.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

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

Xie, J.; Cai, Y.; Chen, J.; Xu, R.; Wang, J.; and Li, Q. 2024a. Knowledge-Augmented Visual Question Answering With Natural Language Explanation. *IEEE Transactions on Image Processing*.

Xie, J.; Cai, Y.; Huang, Q.; and Wang, T. 2021. Multiple Objects-Aware Visual Question Generation. In *MM*, 4546–4554.

Xie, J.; Chen, J.; Fang, W.; Cai, Y.; and Li, Q. 2023. Visual question generation for explicit questioning purposes based on target objects. *Neural Networks*, 167: 638–647.

Xie, J.; Chen, J.; Liu, Z.; Cai, Y.; Huang, Q.; and Li, Q. 2024b. Video Question Generation for Dynamic Changes. *IEEE Transactions on Circuits and Systems for Video Technology*.

Xie, J.; Fang, W.; Cai, Y.; Huang, Q.; and Li, Q. 2022. Knowledge-Based Visual Question Generation. *IEEE Trans. Circuits Syst. Video Technol.*, 32(11): 7547–7558.

Xie, J.; Zhou, Z.; Wu, Z.; Zhang, X.; Wang, J.; Cai, Y.; and Li, Q. 2024c. Automated Defect Report Generation for Enhanced Industrial Quality Control. In *Proc. of AAAI*, volume 38, 19306–19314.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. 2017. Topic Aware Neural Response Generation. In *AAAI*, 3351–3357.

Xu, X.; Wang, T.; Yang, Y.; Hanjalic, A.; and Shen, H. T. 2021. Radial Graph Convolutional Network for Visual Question Generation. *IEEE Trans. Neural Networks Learn. Syst.*, 32(4): 1654–1667.

Xu, X.; Wang, T.; Yang, Y.; Hanjalic, A.; and Shen, H. T. 2021. Radial Graph Convolutional Network for Visual Question Generation. *IEEE Transactions on Neural Networks and Learning Systems*, 1654–1667.

Zhang, W.; Wang, H.; and Zhang, F. 2024. Skip-Timeformer: Skip-Time Interaction Transformer for Long Sequence Time-Series Forecasting. In *Proc. of IJCAI*, 5499–5507.

Zhang, Y.; Hare, J. S.; and Prügel-Bennett, A. 2018. Learning to Count Objects in Natural Images for Visual Question Answering. In *ICLR*.