Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 3, June 2026, pp. 742~752
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i3.pp742-752

Improved interactivity and automated response for visual question answering

Nguyen Ha Manh Khang, Nguyen Tuan Anh, Nguyen Minh Hoang, Bui Thanh Hung
Data Science Laboratory, Faculty of Information Technology, Industrial University of Ho Chi Minh city, Ho Chi Minh city, Vietnam

Article Info

Article history:

Received Oct 24, 2025
Revised Mar 13, 2026
Accepted May 26, 2026

Keywords:

Automatic response
Interaction
Model evaluation
Natural language processing
Visual question answering

ABSTRACT

Visual question answering (VQA) systems have made substantial progress, yet they still face limitations in handling complex or ambiguous queries and supporting real-time interaction due to reliance on large, computationally expensive models that increase latency and restrict practical deployment, particularly in educational contexts. This study aims to develop an efficient and interactive VQA system that enhances answer accuracy while enabling natural two-way communication with users. To achieve this goal, we propose a lightweight multimodal framework based on pre-trained vision–language models such as BLIP and fine-tuning T5, combined with prompt engineering to improve question understanding and answer generation. The system further incorporates conversational context memory and a feedback mechanism that generates clarification questions when user inputs are ambiguous, thereby strengthening interaction capabilities. Experiments are conducted on public benchmark dataset Flickr8k, using single-GPU computational settings to evaluate accuracy, response latency, and interaction effectiveness. The experimental results demonstrate that the proposed approach achieves competitive or superior accuracy compared to heavier baseline models, while significantly reducing inference time and enabling real-time interaction. The main contributions of this work include a lightweight, prompt-driven VQA architecture, an interactive strategy for resolving ambiguous queries, and empirical evidence that efficient models can support accurate and conversational VQA for education and other real-world applications.

Corresponding Author:

Bui Thanh Hung
Data Science Laboratory, Faculty of Information Technology
Industrial University of Ho Chi Minh city
Ho Chi Minh city, Vietnam
Email: buithanhhung@iuh.edu.vn

1. INTRODUCTION

In the context of the rapid and ongoing advancement of artificial intelligence (AI), visual question answering (VQA) has emerged as one of the most promising and highly interdisciplinary research areas. As a multimodal paradigm, VQA combines computer vision, natural language processing, and deep learning techniques to enable machines to perceive visual scenes, reason about their content, and generate meaningful answers to questions expressed in natural language. This capability represents an important step toward narrowing the gap between human cognitive understanding and machine visual perception. VQA systems are designed not only to recognize objects, attributes, and relationships within images, but also to integrate these visual cues with linguistic comprehension in order to produce coherent and contextually appropriate responses [1]–[5].

The practical significance of VQA technology has been increasingly recognized across various domains, including digital education, intelligent tutoring systems, healthcare analysis, virtual assistants, and visual learning environments [6]–[9]. In education, VQA enables interactive learning by allowing students to ask questions about images, diagrams, or infographics and receive immediate feedback. Similarly, in assistive technologies, VQA can support visually impaired individuals by describing visual scenes or answering questions about their surroundings. These applications demonstrate the potential of VQA to improve accessibility, enhance user engagement, and provide fast, context-aware responses that approximate human-like understanding [10]–[12].

Despite these advantages, current VQA systems still face several challenges that limit their real-world deployment. Their performance often declines when dealing with complex or ambiguous questions that require deeper reasoning, multi-step inference, or external knowledge. In addition, real-time interaction remains difficult due to the heavy computational requirements of large neural architectures, particularly transformer-based vision–language models. Issues related to scalability and resource consumption, such as high GPU memory usage and long inference times, further restrict the use of VQA in low-latency or resource-constrained environments. These limitations reduce system flexibility and hinder applications that require robust and immediate responses [2].

Addressing these issues requires more efficient model architectures, improved multimodal representation learning, and stronger reasoning mechanisms that balance accuracy and computational efficiency. Overcoming these challenges could enable future VQA systems to function as intelligent agents capable of supporting real-time, human-like interaction across diverse visual and linguistic contexts [4], [5].

Early benchmarks, such as the VQA dataset introduced by Antol *et al.* [13], provided a foundation for systematic evaluation but revealed strong dataset biases that encouraged shallow reasoning. Later approaches incorporated attention-based, co-attention, and transformer-based models to improve multimodal alignment, though often at the cost of increased complexity. While large pre-trained vision–language models achieve strong accuracy, prior studies note limited focus on real-time interaction, dialogue-based reasoning, and clarification mechanisms. As a result, most existing systems still operate in a single-turn answering paradigm without conversational feedback [14], [15].

To address these gaps, this study proposes a lightweight, prompt-driven VQA framework that prioritizes efficiency and interactivity. The approach leverages a compact BLIP-based vision–language model combined with prompt engineering to improve question understanding and answer generation while reducing computational overhead. In addition, the system integrates conversational context memory and an active clarification mechanism, enabling it to ask follow-up questions when user inputs are ambiguous. Unlike prior static VQA systems, the proposed framework supports real-time, two-way interaction and adaptive reasoning, making it particularly suitable for educational and interactive applications.

Our approach includes: using the T5TP3 model to generate questions from photo captions; applying a compact BLIP model combined with prompt engineering techniques to optimize input queries and generate descriptive and information-rich answers; build a descriptive VQA dataset from Flickr8k and design a resource-optimized training process. Key contributions to the study include:

- Automatically generate questions using T5TP3 from image captions, producing diverse and contextually appropriate questions for visual content.
- Integrate lightweight BLIP with prompt engineering to improve the accuracy, coherence, and descriptive quality of generated answers.
- Build a descriptive VQA dataset from Flickr8k, where answers correspond to full captions, enabling richer responses than traditional VQA datasets.
- Use a resource-efficient training and evaluation framework, including freezing the vision encoder, gradient accumulation, fp16 optimization, and metrics such as mean question similarity, mean question–caption similarity, unique question ratio, BLEU, and ROUGE.

In addition to the introduction, the remainder of this paper is organized as follows. Part 2 presents the proposed model and provides a detailed analysis of its individual components. Part 3 describes the experimental setup and reports a comparative evaluation of our approach against existing methods. Finally, Part 4 summarizes the main conclusions of this study and discusses potential directions for future research.

2. METHOD

2.1. The proposed method

The proposed system is designed as a two-stage framework that integrates both automatic question generation and intelligent answer prediction, enabling a more seamless and contextually grounded interaction between visual understanding and language reasoning. Specifically, the framework operates through two primary phases: i) Automatic question generation from images, where visual inputs are processed to produce

linguistically diverse and semantically coherent questions that reflect the key elements, objects, and relationships within the image; and ii) Question answering based on image content, in which the generated or user-provided question is analyzed and answered using a fine-tuned BLIP model [16] in conjunction with prompt engineering to ensure contextual accuracy and human-like fluency.

Figure 1 presents the overall workflow of the proposed method, which involves several interconnected steps that bridge visual perception and natural language reasoning. The system first extracts high-level visual representations from the input image, then leverages the T5TP3 [17] based question generation module to automatically formulate relevant questions. Subsequently, the fine-tuned BLIP model, guided by carefully crafted prompts, interprets both the image features and the textual query to generate an answer that aligns with the visual context. This pipeline not only enhances the depth and diversity of question–answer pairs but also improves system adaptability across different domains and interaction scenarios. The detailed procedure of each stage is described as follows.

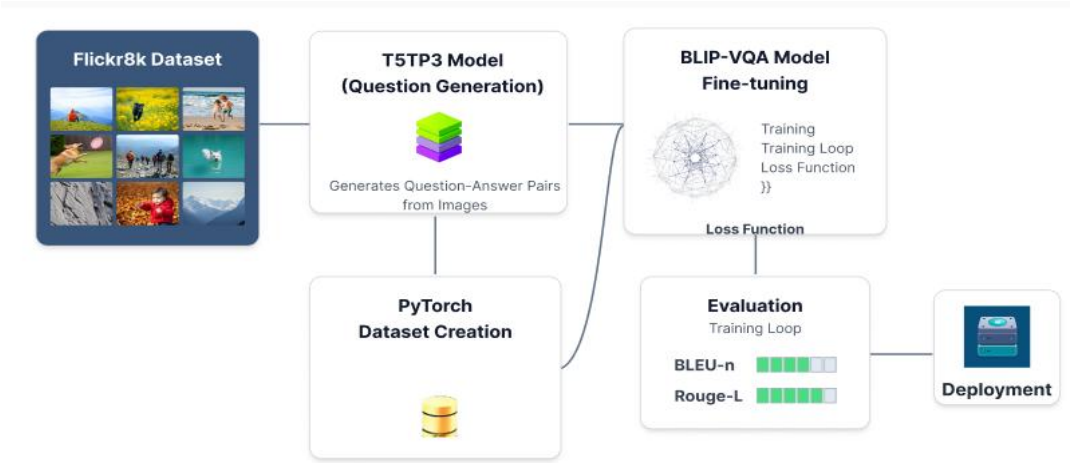

Figure 1. Overview architecture of the proposed model

Data pre-processing: Images and captions from the Flickr8k dataset are used as input. Captions are standardized, and images are converted into a format suitable for the visual model. Automatic question generation using T5TP3: The T5TP3 model takes image captions as input and generates relevant questions, creating contextual question–answer pairs that reflect the image content. Descriptive VQA dataset construction: Images, generated questions, and original captions are combined to form training samples (image, question, answer). Answers are maintained as detailed descriptions rather than the short responses typical of traditional VQA datasets. BLIP-VQA model training: BLIP-VQA is used as the question-answering model. The Vision Encoder is frozen, while gradient accumulation and FP16 are applied to reduce computational cost. Prompt engineering is used during training and inference to encourage coherent and complete answers. Performance evaluation: Question quality is assessed using Mean Question Similarity, Mean Question–Caption Similarity, and Unique Question Ratio. Answer quality is evaluated using BLEU-n and ROUGE-L, enabling a comprehensive assessment of both the dataset and model performance. The above process ensures that the model is both capable of automatically generating contextual VQA training data, optimizing training and inference for a resource-constrained environment, and providing answers that are descriptive and close to natural language. We will describe each part in detail in the next session.

2.2. Generate question-answer pairs from images

The model proposed in this study is designed to enhance both the quality and contextual relevance of questions and answers within a VQA framework. Unlike conventional systems that rely solely on pre-defined question–answer pairs, the proposed approach integrates three complementary components the T5TP3 question generation model, a fine-tuned BLIP model, and prompt engineering techniques to form a cohesive and adaptive architecture.

Through this integration, the system generates semantically rich questions from visual inputs and produces accurate, context-aware, human-like answers by leveraging multimodal alignment between textual and visual features. The T5TP3 component ensures linguistic diversity and grammatical fluency in generated questions, while the fine-tuned BLIP model strengthens visual–textual reasoning and semantic coherence

between image understanding and language output. Prompt engineering further refines model behavior, enabling the system to adapt to different contexts, question types, and response styles without extensive retraining. The main features of this module include:

Automatic question generation from images:

- The T5TP3 model generates questions based on image captions from the Flickr8k dataset.
- Generated questions provide diverse and meaningful contexts that help the model utilize image information effectively.

Context-rich descriptive VQA dataset:

- Answers are detailed descriptive captions rather than short responses, allowing the system to produce more informative outputs.
- Data is preprocessed and organized as a PyTorch dataset to facilitate training.

Efficient BLIP-VQA training:

- The vision encoder is frozen to reduce computational overhead.
- Gradient accumulation and FP16 are applied to optimize memory usage and training time.
- Prompt engineering is incorporated to improve answer coherence and quality.

Multi-criteria evaluation:

- Question quality is assessed using mean question similarity, mean question–caption similarity, and unique question ratio.
- Answer quality is evaluated using BLEU-n and ROUGE-L to measure both accuracy and completeness.

This approach enables training and deployment in resource-limited environments while maintaining the ability to generate descriptive questions and answers, making it suitable for applications such as learning assistants and visual accessibility systems.

The data flow is illustrated in Figure 2 and includes the following steps:

- Image and caption → question generation (T5TP3).
- Construction of a descriptive VQA dataset (image, question, answer).
- Fine-tuning BLIP-VQA with prompt engineering.
- Input: image (ViT) and tokenized question.
- Output: descriptive answer.

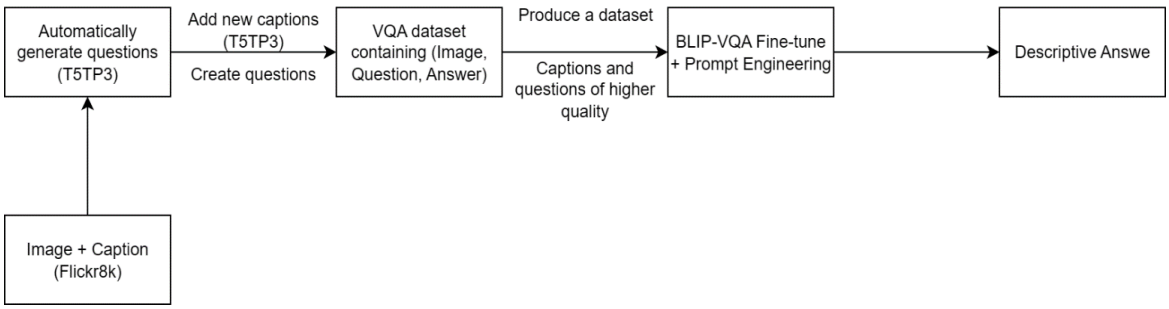

Figure 2. Data flow description

Prompting techniques: During the training and assessment process, we followed prompt methods presented in [18], [19] to do prompting techniques to improve the quality of questions and answers. Question generation was performed using the T5TP3 model fine-tuned on the valhalla/t5-base-qg-hl model, with the input prompt being captions describing images from the Flickr8k episode. We use prompts for both training and testing. The prompt looks like "You are a descriptive VQA assistant. Question: {q}". The question will be reformatted with the prompt. For example, Question: "Who is going into a wooden building?" will result in "You are a descriptive VQA assistant. Question: Who is going into a wooden building?". The automatically labeled questions and answers are then used to train the BLIP model in an image → processor (image, question) → answer model.

Thanks to the above improvements, the system not only increases accuracy but also improves the quality of the user experience through natural language interaction, informative descriptions, and flexible responses. This is a step away from the traditional VQA model to a more descriptive and humane visual Q&A system.

2.3. Model fine-tuning

In this section, we describe the process of answering questions from image content using the fine-tuned BLIP model with prompt engineering. The system uses BLIP-VQA as the core module to fuse visual and textual information. BLIP is chosen for its compact yet effective end-to-end architecture, enabling image–text alignment, feature extraction, and answer generation in a unified model. With prompt engineering, the model better adapts to different question types and produces more descriptive and contextually relevant answers.

The fine-tuned BLIP model also incorporates attention mechanisms inspired by hierarchical co-attention and the stacked attention network (SAN) to strengthen interactions between visual regions and text representations. These mechanisms iteratively align image features with linguistic tokens, improving the model's ability to capture spatial dependencies, object relationships, and cross-modal semantics, leading to more accurate and grounded answers.

However, traditional attention-based approaches often rely on localized visual features and show limitations in multi-step reasoning, long-range dependency modeling, and deeper contextual inference. To address this, this study integrates the BLIP framework with prompt engineering to guide the model's reasoning process. By dynamically refining prompts during inference, the system focuses on informative visual regions and linguistic cues, reducing ambiguity and improving compositional reasoning. This strategy also enables more detailed and context-aware descriptions for complex scenes. Overall, the approach balances computational efficiency, interpretability, and semantic expressiveness, allowing the VQA system to produce accurate, coherent, and informative answers.

In the freeze vision encoder, we apply gradient accumulation and FP16 to reduce computation costs and prompt engineering in training and reasoning to guide the model to generate complete and coherent answers. Figure 3 describes a detailed description of the fine-tuning process of BLIP. This processing includes the following steps:

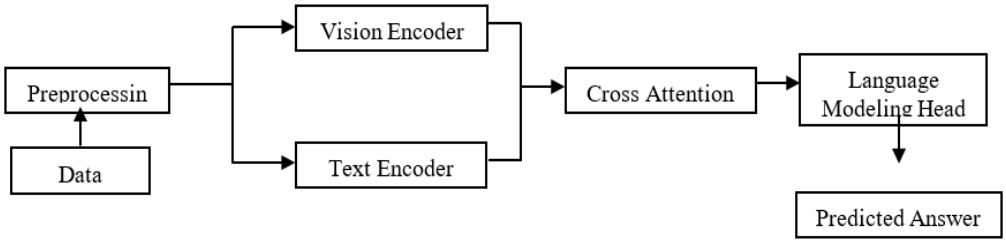

Figure 3. Detailed description of the fine-tuning process of BLIP

Forward pass: When using HuggingFace's Trainer to fine-tune BLIP for the VQA task, the forward pass proceeds as follows. Data input: Images are loaded from the DataLoader and converted into pixel_values tensors according to BLIP requirements, while questions are tokenized into input_ids and attention_masks using the BLIP tokenizer. Image feature extraction: The vision encoder (typically a Vision Transformer – ViT) processes the image tensors to produce embeddings that capture high-level visual features such as objects and colors. Question feature extraction: The Transformer-based text encoder processes the tokenized question and generates text embeddings. Multimodal fusion: Cross-attention layers enable interactions between textual and visual representations, producing a contextual multimodal representation. Answer prediction: The fused representation is passed to a language modeling head (linear layer with softmax) to compute token probabilities, from which the final answer is generated.

Loss calculation: The trainer calls the compute_loss function in BlipForQuestionAnswering. Predictions are compared with ground-truth answers at the token level using cross-entropy loss. Padding tokens are masked so that loss is computed only on valid tokens. The resulting loss reflects the difference between predicted and reference answers for the current batch.

Backpropagation: During the backward pass, PyTorch computes gradients of the loss for parameters in the vision encoder, text encoder, and cross-attention layers. The AdamW optimizer updates the weights to minimize the loss. Key hyperparameters include the learning rate, which controls update magnitude, and weight decay, which helps prevent overfitting. A learning-rate scheduler such as linear Warmup and Decay is typically applied to gradually increase the learning rate at the start of training and decrease it later to improve convergence.

2.4. Evaluation

We use evaluation indicators that are suitable for the problem of generating automatic answers for VQA: BLEU [20] and ROUGE-L [21]. BLEU-1, BLEU-2, BLEU-3, BLEU-4: N-gram-based measurements to assess the similarity between the resulting answer and the actual answer. BLEU-1: Evaluation of unigram duplication, BLEU-2: Duplicate assessment of 2 consecutive words (bigram), BLEU-3, BLEU-4: Expands with 3- and 4-word phrases (trigram, 4-gram) that help reflect the coherence of the resulting answer. BLEU metric is calculated in (1):

$$BLEU = BP \times exp(\sum_{n=1}^{N} w_n \times \log p_n) \quad (1)$$

where: N-gram matching: Counts the number of n-grams that coincide with the reference sentence; Precision with adjustment: Calculates the n-gram match ratio and applies clipping to avoid repeating the word fraud; and brevity penalty (BP): Penalty when the birth sentence is too short for the reference sentence. ROUGE-L measures the similarity between the generated and reference texts, using the longest common subsequence (LCS) to assess content matching without contiguity. This metric is based on the recall and precision of LCS to measure the match between the birth sentence and the reference. In addition, we use the following measurements to evaluate the question generated:

Mean question similarity (MQS): Average similarity (e.g., cosine similarity) among generated questions in the dataset, measuring question diversity and relevance. Mean question–caption similarity (MQCS): Average similarity between each generated question and its original caption, evaluating how well questions reflect caption content. Unique question ratio (UQR): Percentage of unique (non-repeated) questions among all generated questions, indicating the diversity of the question generation system.

3. RESULTS AND DISCUSSION

3.1. Dataset

The dataset used in the experiment was Flickr8k [22], which consisted of 8000 images depicting situations in everyday life. Each photo has several short captions in English. Table 1 shows dataset statistics. To build a VQA dataset from an image, we proceed with the following steps: Caption preprocessing: duplicate removal, standardize text; Create a question from the caption using the T5TP3 question generation model; Assign the answer to the corresponding caption itself (equivalent to the descriptive VQA model) and each template includes: image, question, answer, and is saved to a file that makes up the fine-tuned Flickr8k dataset.

Table 1. Dataset statistics

Dataset	Number
Train set	5663
Test set	2428

3.2. Result

We used Kaggle Notebooks, a cloud-based platform provided by Kaggle. The computing environment is as follows: CPU: Intel(R) Xeon(R) CPU @ 2.00GHz; GPU: NVIDIA Tesla P100-PCIE-16GB (VRAM: 16,384 MiB ≈ 16 GB); Operating System: Ubuntu 22.04.4 LTS; Python Version: Python 3.11.13; Key Libraries: BLIP, T5, PIL, spacy, pandas, scikit-learn, tqdm, and gc are provided within the notebook. We evaluated question generation results using MQCS, MQS, and UQR metrics by comparing with BART-BASE and FLAN-T5, the results are© shown in Table 2.

Table 2. Comparison of question generation results

Metric	BART-BASE	T5TP3	FLAN-T5
MQCS	0.9527	0.6802	0.4212
MQS (With 1000 Sample)	0.1599	0.2572	0.3522
UQR	0.9937	0.9327	0.3657

The evaluation of T5TP3 across three metrics demonstrates its overall effectiveness for caption-to-question generation. Specifically, T5TP3 achieves an MQCS score of 0.6802, which can be considered moderate. This result indicates that the generated questions do not strictly follow the caption wording but instead tend to expand upon the original content by introducing additional semantic elements. The MQS score of 0.2572 reflects a moderate level of repetition in question types, which can be attributed to the

model's tendency to learn common interrogative patterns (e.g., what, why, how). Meanwhile, T5TP3 attains a high UQR score of 0.9327, indicating strong semantic diversity, as the model is capable of generating questions from multiple perspectives, such as spatial relationships, purpose, and contextual information. Overall, these evaluation results suggest that T5TP3 is the most suitable model for the task of generating questions from captions. Compared to BART-BASE, although T5TP3 yields a slightly lower MQCS score, this characteristic reflects its strength in not only reformulating captions but also expanding their semantic content to produce more informative and exploratory questions. At the same time, the consistently high UQR score demonstrates that the model maintains question diversity without sacrificing relevance to the caption. In contrast, BART-BASE primarily emphasizes surface-level form variation, while FLAN-T5 tends to generate more generalized and repetitive questions. Therefore, considering the balance among relevance (MQCS), semantic diversity, and question uniqueness (UQR), T5TP3 was selected as the primary model for this study.

In this study, we fine-tune the BLIP-VQA (Salesforce/blip-vqa-base) model on the Flickr8k dataset reprocessed as VQA. We use the T5TP3 auto-question generation model to generate questions from the original caption, and then fine-tune the BLIP model to generate descriptive answers. The results were evaluated using BLEU and ROUGE, showing that the proposed model is superior to previous methods such as BLIP-2 [23], InstructBLIP [24], DEiT + Bert [1], BEiT + GPT2 [1]. Table 3 shows comparison results between methods.

After fine-tuning the BLIP-T5TP3 model with the autogenerated data from the Flickr8k set, we evaluated the model on the test set and obtained the following results: BLEU-1 scored 0.32, BLEU-2 scored 0.27, BLEU-3 scored 0.23, BLEU-4 scored 0.19, and ROUGE-L scored 0.52. These results are relatively higher than the previous two models, as the high ROUGE results prove that the model's answers retain meaning and cover the content well. The uniform increase in BLEU at n-grams means that the model not only matches the idea but also does so more accurately in structure and vocabulary. This suggests that the fine-tuned BLIP model is more likely to produce more semantic, descriptive answers than the original models that have not been fine-tuned.

Figure 4 illustrates the distribution of BLEU-n scores (n = 1 to 4) obtained by the BLIP model after fine-tuning on the constructed VQA dataset. As shown in the figure, the BLEU-1 scores exhibit the highest values and are predominantly concentrated in the range of 0.5–0.7, indicating that the model is highly effective at generating accurate single-word or unigram-level answers that closely match the reference responses. This suggests strong performance in recognizing key visual entities, attributes, and simple concepts within images. In contrast, BLEU-2 and BLEU-3 scores show a noticeable decline, reflecting a moderate ability to capture short-range contextual relationships, such as bigrams and trigrams, where limited compositional structure is required. This decline implies that while the model can produce locally coherent phrases, its consistency decreases as contextual dependencies increase.

Notably, the BLEU-4 scores are strongly skewed toward lower values, with most samples in the 0.1–0.3 range. This distribution indicates that the model still struggles to generate longer sequences that closely match ground-truth answers at the four-gram level. In many cases, generated responses align only with partial segments of the reference sentence, reducing precision for complex grammatical structures and semantically rich descriptions. This behavior suggests a tendency toward simplification when handling longer or more complex answers. However, this pattern is common in automated text generation and VQA systems, where models often prioritize semantic correctness and visual relevance over exact lexical matching. Therefore, lower BLEU-4 scores do not necessarily indicate poor answer quality but rather reflect the trade-off between fluency, semantic adequacy, and strict n-gram overlap in generative evaluation metrics.

Figure 5 shows the distribution of ROUGE-L scores on the test set, which follows an approximately normal pattern with a slight skew toward lower values. Most samples fall between 0.4 and 0.65, indicating that the model often generates answers with considerable overlap and structural similarity to the reference responses. The average score of about 0.52 (marked by the red line) suggests a reasonably acceptable performance level for generative VQA tasks and reflects the model's ability to preserve key semantic information and sentence structure.

Table 3. Comparison results between methods

Method	BLEU @1	BLEU @2	BLEU @3	BLEU @4	ROUGE
BLIP-2	0.22	0.19	0.16	0.14	0.42
InstructBLIP	0.08	0.06	0.05	0.04	0.15
BLIP-T5TP3 (ours)	0.32	0.27	0.23	0.19	0.52
DEiT + Bert	0.17	0.10	0.07	0.05	0.26
BEiT + GPT2	0.24	0.13	0.08	0.05	0.25

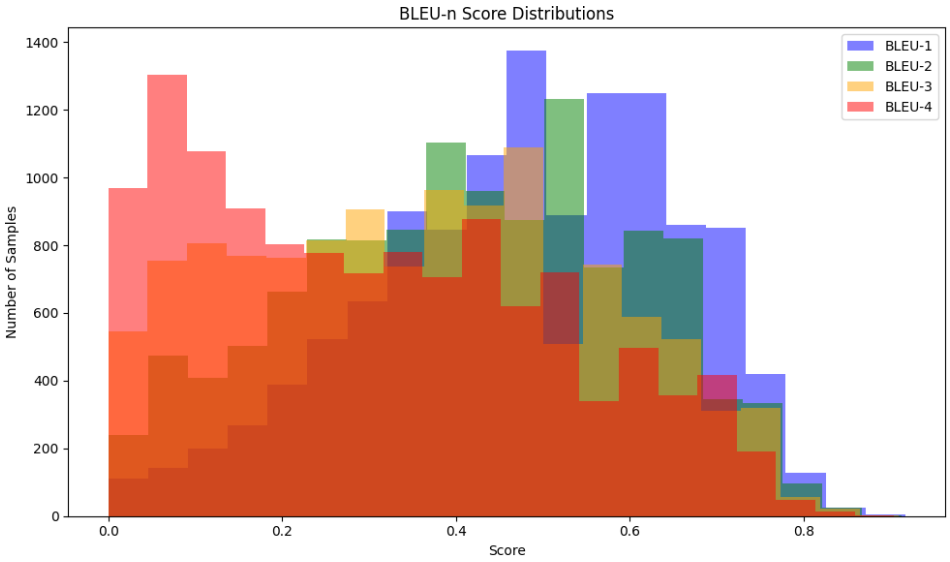

Figure 4. BLEU-1 to BLEU-4 score distribution chart on the test dataset

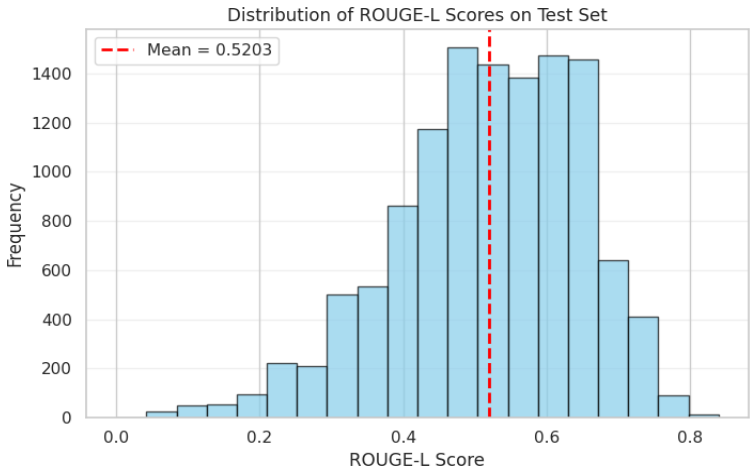

Figure 5. ROUGE-L distribution chart on test set

However, the distribution remains relatively wide, with some samples scoring below 0.2 and others approaching 0.8. This variability indicates inconsistent performance across question types and visual contexts. The model performs well on straightforward or visually grounded questions but struggles with more complex or ambiguous queries. High-scoring cases confirm the model's capability to generate responses closely matching the references, while low-scoring outliers reveal limitations in robustness and generalization.

Overall, the results indicate promising answer generation ability, though further improvements are needed to reduce performance variance and increase consistency across diverse visual and linguistic scenarios. Response time comparisons between methods are presented in Table 4. The response time reported in Table 4 reveals an apparent paradox. Although the BLIP-Finetune model (ViT-B/16 combined with a BLIP Decoder) is considerably smaller in terms of model parameters than InstructBLIP (ViT-g/14 paired with FLAN-T5-XL), it nevertheless exhibits a significantly higher average response time (0.5516s compared to 0.0907s). This discrepancy cannot be attributed to raw computational complexity or model size alone. Instead, the primary technical cause lies in the generation and decoding strategy employed during inference. Specifically, the BLIP-T5TP3 model frequently generates longer output sequences due to unresolved word repetition issues. As a result, its generated responses often reach the predefined max_length threshold (e.g., 64 or even 128 tokens), leading to extended decoding loops and increased latency. In addition, the stopping mechanism in BLIP-Finetune is less effective, as the model does not consistently predict the end-of-sequence (<EOS>) token at an early stage. In contrast, InstructBLIP and BLIP-2 typically produce much shorter and

more concise responses, often containing fewer than 10 tokens. These models tend to predict the <EOS> token very early in the decoding process, allowing generation to terminate promptly. Consequently, despite their larger parameter sizes, their inference time remains substantially lower due to reduced token-by-token decoding overhead.

Table 4. Response time results between methods

Method	Backbone	AVG Response time (s)
BLIP-2	ViT-g/14 + FLAN-T5-XL	0.2189
InstructBLIP	ViT-g/14 + FLAN-T5-XL	0.0907
BLIP-T5TP3 (ours)	ViT-B/16 + BLIP Decoder	0.5516
DEiT + Bert	DEiT-B/16 + BERT-base	0.0351
BEiT + GPT2	ViT-B/16 + gpt2	0.1781

We analyzed all incorrect predictions of the testset and saw several inherent limitations of the BLIP-based VQA model. First, the model struggles with questions that require background or world knowledge beyond what is explicitly visible in the image, such as identifying the name of the "Oklahoma University" football team. This limitation arises because the current system does not incorporate external knowledge retrieval mechanisms or text recognition (OCR) modules that could provide additional semantic information. Second, the model demonstrates limited capability in quantitative reasoning and fine-grained semantic understanding, as evidenced by errors in counting the number of people in an image or recognizing complex emotional states such as "excited." These tasks require not only object detection but also contextual interpretation of facial expressions, body language, and group dynamics, which are not explicitly modeled. Finally, the system is prone to object and action confusion, for instance mistaking a "cannon" for a "tire swing," indicating weaknesses in recognizing specialized objects and inferring collective or event-level actions. Overall, these errors highlight that the model relies primarily on visual features and question prompts, without support from external knowledge sources or specialized reasoning modules, which limits its robustness in complex real-world scenarios.

4. CONCLUSION

This study makes several measurable contributions to VQA research by improving interoperability and automatic answer generation through the integration of a BLIP-T5 VQA model with an automated question generation pipeline. A new VQA dataset was constructed from the Flickr8k dataset, where image captions were automatically transformed into question–answer pairs using the T5TP3 model, enabling scalable data creation without manual annotation. Fine-tuning BLIP-T5 VQA on this dataset produced consistent improvements over baseline models such as BLIP-2, InstructBLIP, DEiT + Bert, and BEiT + GPT2 according to BLEU-1 to BLEU-4 and ROUGE-L metrics. Experimental results show particularly strong gains in BLEU-1 and BLEU-2, reflecting accurate recognition of key visual concepts and short answers, while lower BLEU-3 and BLEU-4 scores indicate ongoing challenges in generating longer, compositional responses. ROUGE-L results demonstrate moderate structural similarity to reference answers, though with some variance across samples.

Despite these strengths, the approach still has limitations. Performance may decline on complex reasoning tasks requiring extensive external knowledge or multi-step inference, and the clarification mechanism relies on predefined prompting heuristics rather than adaptive dialogue policies. In addition, evaluations were conducted mainly on benchmark datasets and limited real-world scenarios, which may not fully capture diverse user behavior. Future research should investigate adaptive prompt learning, parameter-efficient fine-tuning, multi-turn dialogue modeling, and large-scale user-in-the-loop evaluations to further improve robustness, generalization, and practical applicability.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nguyen Ha Manh Khang	✓	✓	✓	✓		✓		✓	✓	✓				
Nguyen Tuan Anh				✓			✓			✓	✓			
Nguyen Minh Hoang			✓				✓			✓	✓			
Bui Thanh Hung	✓	✓			✓	✓		✓	✓	✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in [University of Illinois Urbana-Champaign] at https://forms.illinois.edu/sec/1713398, reference number [22].

REFERENCES

[1] B. T. Hung and H. V. H. Duy, "ExVQA: a novel stacked attention networks with extended long short-term memory model for visual question answering," *Computers and Electrical Engineering*, vol. 126, p. 110439, Aug. 2025, doi: 10.1016/j.compeleceng.2025.110439.

[2] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu, and W. Zheng, "The multi-modal fusion in visual question answering: a review of attention mechanisms," *PeerJ Computer Science*, vol. 9, p. e1400, May 2023, doi: 10.7717/peerj-cs.1400.

[3] S. Yang, C. Han, S. Luo, and E. Hovy, "MAGIC-VQA: Multimodal and grounded inference with commonsense knowledge for visual question answering," in *Findings of the Association for Computational Linguistics: ACL 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 16967–16986. doi: 10.18653/v1/2025.findings-acl.872.

[4] A. Pandey, D. Bodo, A. Phukan, and A. Ekbal, "The quest for visual understanding: a journey through the evolution of visual question answering." Jan. 13, 2025. [Online]. Available: http://arxiv.org/abs/2501.07109

[5] H. J. Singh, G. Bathla, M. Mehta, G. Chhabra, and P. Singh, "Visual questions answering developments, applications, datasets and opportunities: A state-of-the-art survey," in *2nd International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2023 - Proceedings*, IEEE, Mar. 2023, pp. 778–785. doi: 10.1109/ICSCDS56580.2023.10104870.

[6] B. T. Hung, "Content-based image retrieval using multi-deep learning models," in *Lecture Notes in Networks and Systems*, vol. 445, 2023, pp. 347–357. doi: 10.1007/978-981-19-1412-6_29.

[7] B. T. Hung, "Link prediction in paper citation network based on deep graph convolutional neural network," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 117, 2022, pp. 897–907. doi: 10.1007/978-981-19-0898-9_67.

[8] B. T. Hung and V. Q. Huy, "MTFIC: enhanced fashion image captioning via multi-transformer architecture with contrastive and bidirectional encodings," *Visual Computer*, vol. 41, no. 13, pp. 10841–10855, Oct. 2025, doi: 10.1007/s00371-025-04072-8.

[9] B. T. Hung, N. V. P. Nhan, and N. T. Sy, "DCARES: deep convolutional neural network with neural-based optimization for image-based product recommender system," *Multimedia Tools and Applications*, vol. 84, no. 30, pp. 36693–36723, Feb. 2025, doi: 10.1007/s11042-025-20655-y.

[10] S. Chowdhury and B. Soni, "R-VQA: A robust visual question answering model," *Knowledge-Based Systems*, vol. 309, p. 112827, Jan. 2025, doi: 10.1016/j.knosys.2024.112827.

[11] N. D. Huynh, M. R. Bouadjenek, S. Aryal, I. Razzak, and H. Hacid, "Visual question answering: from early developments to recent advances -- a survey." Jan. 11, 2025. [Online]. Available: http://arxiv.org/abs/2501.03939

[12] M. Yamada, V. D'amario, K. Takemoto, X. Boix, and T. Sasaki, "Transformer module networks for systematic generalization in visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10096–10105, Mar. 2024, doi: 10.1109/TPAMI.2024.3438887.

[13] S. Antol *et al.*, "Vqa: visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[14] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: a survey of methods and datasets," *Computer Vision and Image Understanding*, vol. 163, pp. 21–40, 2017, doi: 10.1016/j.cviu.2017.05.001.

[15] R. Kabir, N. Haque, M. S. Islam, and Marium-E-Jannat, "A comprehensive survey on visual question answering datasets and algorithms." Nov. 17, 2024. [Online]. Available: http://arxiv.org/abs/2411.11150

[16] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," *Proceedings of Machine Learning Research*, vol. 162, pp. 12888–12900, 2022.

[17] C. Zhang, H. Zhang, Y. Sun, and J. Wang, "Downstream transformer generation of question-answer pairs with preprocessing and postprocessing pipelines," in *DocEng 2022 - Proceedings of the 2022 ACM Symposium on Document Engineering*, 2022. doi: 10.1145/3558100.3563846.

[18] K. Zhu *et al.*, "PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts," in *LAMPS 2024 - Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, New York, NY, USA: ACM, Nov. 2024, pp. 57–68. doi: 10.1145/3689217.3690621.

[19] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: techniques and applications," 2025, [Online]. Available: http://arxiv.org/abs/2402.07927
[20] B. S. U. Kim, K. I. M. Jieun, L. E. E. Deokwoo, and B. Jang, "Visual question answering: a survey of methods, datasets, evaluation, and challenges," *ACM Computing Surveys*, vol. 57, no. 10, 2025, doi: 10.1145/3728635.
[21] J. Ma *et al.*, "Robust visual question answering: datasets, methods, and future challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5575–5594, 2024, doi: 10.1109/TPAMI.2024.3366154.
[22] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8692 LNCS, no. PART 4, 2014, pp. 529–545. doi: 10.1007/978-3-319-10593-2_35.
[23] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of Machine Learning Research*, 2023, pp. 19730–19742.
[24] W. Dai *et al.*, "InstructBLIP: towards general-purpose vision-language models with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36. pp. 49250–49267, 2023.

BIOGRAPHIES OF AUTHORS

Nguyen Ha Manh Khang is a third-year student at the Industrial University of ho Chi Minh City, Vietnam. He is pursuing an Engineer's degree in Information Technology, specializing in Data Science. His research interests include visual question answering and large language model. He can be contacted at email: nghmanhkhang@gmail.com.

Nguyen Tuan Anh is currently a third-year undergraduate student at the Industrial University of Ho Chi Minh City, majoring in Data Science under the Faculty of Information Technology. His main research interests include Visual Question Answering and Image Caption Generation. He can be contacted at email: nguyentuananhck2005@gmail.com.

Nguyen Minh Hoang is currently a third-year student deeply engaged in the Data Science program at the Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam. His main research interests include Visual Question Answering and leveraging computational methods to extract insights from complex datasets. He can be contacted at email: nguyenminhhoangnt20@gmail.com.

Bui Thanh Hung received his M.S. degree and Ph.D. degree from Japan Advanced Institute of Science and Technology, Japan (JAIST) in 2010 and 2013. He has completed 2 projects, published 30 journals, 8 books, 33 book chapters, 48 International conference papers, and 21 domestic conference papers. Now he works for the Data Science Laboratory, the Data Science Department, the Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam. His main research interests are natural language processing, machine learning, machine translation, data science, image processing, voice recognition, and artificial intelligence. He can be contacted at email: buithanhhung@iuh.edu.vn.