Improved interactivity and automated response for visual question answering

Nguyen Ha Manh Khang[1], Nguyen Tuan Anh[1], Nguyen Minh Hoang[1], Bui Thanh Hung[1]
[1]Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city, Ho Chi Minh city, Vietnam

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Automatic Response
Interaction
Natural Language Processing
Model Evaluation
Visual Question Answering

ABSTRACT

Visual Question Answering (VQA) is a significant field in artificial intelligence, integrating computer vision and natural language processing to answer questions based on images. This study focuses on enhancing interaction and improving automatic responses in VQA, aiming to build intelligent systems that effectively meet user demands. VQA not only optimizes user experience but also holds great application potential, particularly in education, where quick and accurate feedback can improve learning outcomes. Current approaches leverage deep learning, combining convolutional neural networks for image analysis and language models for question processing. However, these methods face difficulties in ensuring accuracy for complex questions and often lack flexible real-time interaction. These limitations stem from the use of large models that demand high computational resources and result in slow response times, reducing their practicality. To address these challenges, the study proposes a new approach that utilizes lightweight pre-trained models such as BLIP or MiniGPT, in combination with Prompt Engineering to optimize queries and enhance answer accuracy. The system also integrates conversational context memory and a feedback mechanism, allowing it to ask follow-up questions when user queries are unclear-thereby improving real-time interaction. This approach is based on multimodal learning, enabling the system to adapt to complex contexts and enhance user experience in applications such as education, virtual assistants, or visual guidance. The expected outcome is an enhanced VQA system that is not only more accurate but also capable of supporting effective two-way communication. The research aims to deliver an improved VQA tool that boosts both response quality and interaction capabilities, making it a valuable support system in education and other domains. These advancements promise to accelerate the development of conversational artificial intelligence, contributing to the creation of smarter and more efficient human-computer interaction environments.

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 continuous evolution of artificial intelligence (AI), visual question answering (VQA) has emerged as one of the most promising and interdisciplinary research directions. As a multimodal technology, VQA integrates computer vision, natural language processing, and deep learning to enable machines to interpret and reason about visual scenes and subsequently generate meaningful answers to questions posed in natural language. This capability represents a significant step toward bridging the gap between human cognitive understanding and machine perception. VQA systems not only interpret objects, attributes, and relationships within an image but also combine these visual cues with linguistic comprehension to deliver coherent and contextually relevant responses [1-5]

The practical value of VQA technology is increasingly recognized across various domains such as digital education, intelligent tutoring systems, healthcare diagnostics, virtual assistants, and visual learning support environments. In educational settings, for example, VQA can help learners interact with visual materials more effectively by allowing them to ask questions about images, diagrams, or infographics and receive immediate, informative feedback. Similarly, in the context of assistive technologies, VQA can empower visually impaired users by verbally describing scenes or answering queries about their surroundings. These applications demonstrate how VQA can enhance accessibility, improve user engagement, and boost efficiency by offering fast, accurate, and context-aware responses that simulate human-like understanding [6-9]

Despite these advantages, current VQA systems still face a number of persistent challenges that limit their real-world deployment. First, their accuracy often declines significantly when dealing with complex, ambiguous, or context-rich questions that require deep reasoning, multi-step inference, or external knowledge integration. Second, achieving real-time interaction remains difficult due to the heavy computational demands of large-scale neural architectures, such as transformer-based vision-language models. Third, scalability issues and high resource consumption—both in terms of GPU memory and inference time—pose obstacles for deploying VQA in low-latency environments or on resource-constrained devices. These limitations not only reduce user satisfaction but also restrict the adaptability of VQA systems in scenarios that demand immediate feedback, robustness under diverse conditions, and efficient multimodal reasoning.

Ultimately, addressing these challenges requires the development of more efficient model architectures, better multimodal representation learning techniques, and novel reasoning mechanisms that allow VQA systems to balance accuracy, interpretability, and computational efficiency. By overcoming these barriers, future VQA frameworks could evolve into intelligent agents capable of real-time, human-like interaction across a wide range of visual and linguistic contexts. [10-13].

To overcome this, this study proposes an approach that focuses on both aspects: improving the quality of answers and enhancing the interoperability of the VQA system. 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 photo captions, helping to create diverse and suitable question sets for image content.
- Combine lightweight BLIP with Prompt Engineering to enhance the accuracy, coherence, and detailed responsiveness of your answers.
- Build a descriptive VQA dataset from Flickr8k, where the answer is fully descriptive captions, helping the system provide richer information than traditional VQA sets.
- Lightweight and resource-optimized training process, including freezing vision encoder and using gradient accumulation techniques and fp16 to reduce training costs.
- Dual evaluation framework: evaluate questions using Mean Question Similarity, Mean Question–Caption Similarity, Unique Question Ratio, evaluate answers using BLEU, ROUGE, to comprehensively reflect the quality of the dataset.

In addition to the introduction, the rest of the paper will include the following: Part 2 will present relevant research. The general model will be presented and analyzed in detail by the components of the model in Part 3 of the paper. Part 4 will present the experiment and compare the results of our research with other methods. Part 5 will present the conclusions of our research and the direction of future development.

2. THE COMPREHENSIVE THEORETICAL BASIS

Visual Question Answering represents an interdisciplinary research problem that lies at the intersection of computer vision and natural language processing. It challenges models to jointly interpret visual scenes and natural language queries, requiring not only recognition of objects but also reasoning about spatial, semantic, and contextual relationships. The task was first formally introduced by Antol et al., who proposed the inaugural VQA dataset along with a standardized set of evaluation metrics, thereby establishing a foundational benchmark for subsequent research and model comparison within the community [14]. This pioneering work catalyzed widespread interest in the field, yet it soon revealed several limitations. Specifically, the original dataset exhibited strong biases, enabling models to achieve high accuracy by exploiting superficial statistical patterns or performing shallow visual recognition, rather than demonstrating true reasoning ability. Such observations motivated later studies to design datasets and models that focus more deeply on contextual understanding, compositional reasoning, and cross-modal inference, thus moving beyond simple pattern matching toward more cognitively inspired problem-solving.

To capture the rapid evolution of VQA research, Faria et al. conducted a comprehensive survey summarizing key advances and categorizing methods according to their architectural paradigms, such as attention-based frameworks, co-attention mechanisms, and transformer-based models [15]. Their study systematically compared the performance of these models across multiple benchmark datasets (e.g., VQA v2, CLEVR, GQA), providing an essential overview of how representation learning, pre-training, and multimodal fusion techniques have shaped the field. Importantly, their analysis highlighted the growing reliance on pre-trained multimodal encoders and self-supervised learning, which have significantly improved performance and generalization. However, this survey also identified research gaps, particularly the lack of focus on real-time interaction, dialogue-based VQA, and re-enquiry mechanisms—areas that this present study seeks to address by enhancing system adaptability and responsiveness in dynamic environments.

Further contributions by Kafle & Kanan and Wu et al. provided critical insights into the data characteristics and question-type distributions commonly found in VQA datasets (such as yes/no questions, counting, and object recognition tasks) [16–17]. Their analyses exposed the persistent issue of dataset bias, where models tend to ignore contextual cues or linguistic subtleties, leading to inflated but misleading accuracy metrics. These findings inspired our work toward building context-rich and diverse generative datasets, where the T5TP3 model is employed to automatically generate questions from image captions. This generative approach introduces greater variation in linguistic structure and visual grounding, thereby enhancing the semantic richness and relevance of the questions to the underlying visual content.

From a technical standpoint, Lu et al. introduced the Hierarchical Co-Attention model, while Yang et al. proposed the Stacked Attention Network (SAN)—both of which significantly improved the alignment between visual and textual modalities through refined attention mechanisms [18–19]. Nevertheless, these approaches remain heavily dependent on regional visual features and often struggle with multi-step reasoning, long-range dependency modeling, and explainability, limiting their scalability to more complex reasoning tasks. Building upon these foundations, Li et al. later explored Visual Question Generation (VQG)—the inverse task of VQA—as a parallel research direction aimed at promoting active human–machine interaction [20]. Although promising, their framework has not yet been extended to real-time or domain-specific VQA applications, such as those involving contextual understanding in dynamic environments (e.g., surveillance or interactive learning systems), leaving substantial potential for further exploration.

Recent advancements in deep learning have profoundly transformed natural language understanding and vision-language integration within VQA. The adoption of deep neural architectures, particularly Convolutional Neural Networks (CNNs) for image feature extraction and Transformer-based models for cross-modal reasoning, has enabled systems to perform fine-grained alignment between visual elements and textual components. This synergy facilitates contextually coherent answer generation and enhances reasoning over complex visual scenes [21–24]. As research continues to evolve, it is evident that the combination of powerful deep learning models, generative data augmentation, and real-time multimodal reasoning will define the next generation of VQA systems—ones that are not only accurate and explainable but also interactive, adaptive, and capable of engaging in meaningful dialogue with users.

In the Vietnamese context, Tran et al. and Nguyen et al. introduce ViVQA and OpenViVQA, respectively, which provide data and a baseline model for the low-resource language [25-26]. However, the scale of data is still limited, and these systems have not integrated a two-way interaction mechanism.

From the above analysis, it can be seen that VQA has made a lot of progress, but there are still gaps: the ability to understand the context with complex questions is not high, real-time interactivity is still limited, large resource requirements hinder practical application, and there is no effective integration between images, language and conversation. This research aims to solve the above problems by using the T5TP3 model to generate questions from captions, applying a lightweight BLIP model combined with Prompt Engineering to optimize queries and improve answer accuracy, towards effective implementation in education, virtual assistants and intelligent human-machine communication systems.

3. METHOD

3.1. 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 [27] in conjunction with Prompt Engineering to ensure contextual accuracy and human-like fluency.

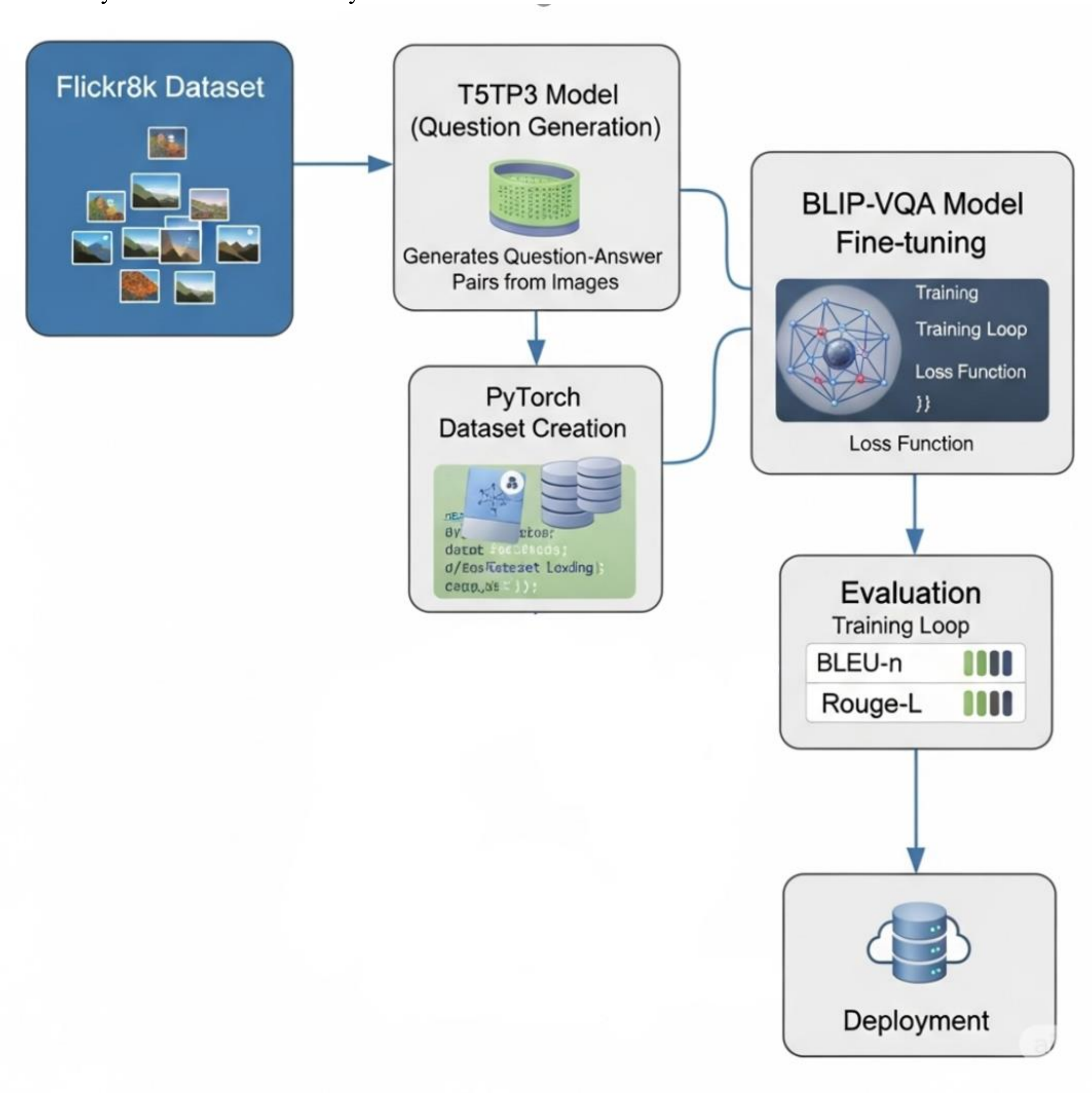


Figure 1. Overview archiecture of the proposed model

As illustrated in Figure 1, the overall workflow of the proposed method 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 [28]-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:

Data pre-processing

- The set of photos and captions from Flickr8k are used as input data sources.
- Captions are standardized text, photos are converted to a format suitable for the visual model.

Automatic question generation using T5TP3

- The T5TP3 model receives the image caption as input and generates relevant questions.
- This process helps to create a set of contextual questions-answers that naturally reflect the content of the photo.

Create a descriptive VQA dataset

- Combine photos, generated questions, and original captions into templates of training data (images, questions, answers).
- The answers are kept in the form of detailed descriptions, which is different from the short answer style in traditional VQA sets.

BLIP-VQA model training

- Use BLIP-VQA as a question answering model.
- Freeze Vision Encoder, apply gradient accumulation and FP16 to reduce computation costs.
- Prompt Engineering is applied in training and reasoning to guide the model to generate complete and coherent answers.

Performance evaluation

- Question Quality: Mean Question Similarity, Mean Question–Caption Similarity, Unique Question Ratio.
- Answer quality: BLEU-n and ROUGE-L.

These criteria allow for a comprehensive evaluation of both aspects: data and models.

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 decribe each part in detail in the next sesion.

3.2. 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 Visual Question Answering 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 not only generates diverse and semantically rich questions from visual inputs but also produces more accurate, context-aware, and human-like answers by leveraging multimodal alignment between textual and visual features. The T5TP3 component ensures the linguistic diversity and grammatical fluency of the generated questions, while the fine-tuned BLIP model strengthens visual-textual reasoning and semantic coherence between image understanding and language output. Meanwhile, Prompt Engineering serves as a control mechanism that refines the model's behavior, allowing the system to dynamically adapt to different contexts, question types, and response styles without extensive retraining. The main features of this module include:

Generate questions automatically from photos

- Use the T5TP3 model to generate questions based on photo captions from Flickr8k episodes.
- The generated question has a clear and diverse context, helping the model to exploit the image information effectively.

Context-rich and descriptive VQA datasets

- The answer is a detailed description caption instead of a short answer, helping the system provide more complete and natural information.
- Data is preprocessed and organized as a PyTorch Dataset for the convenience of training.

Compact BLIP-VQA Training

- Freeze Vision Encoder to reduce computational overhead.
- Apply gradient accumulation and FP16 to optimize memory and training time.
- Incorporate Prompt Engineering to improve the quality and coherence of answers.

Multi-criteria assessment

- Questions are evaluated using Mean Question Similarity, Mean Question–Caption Similarity, and Unique Question Ratio.
- Answers are evaluated using BLEU-n and ROUGE-L to reflect both accuracy and completeness.

This model allows for training and deployment in a limited computational environment, while still ensuring the ability to generate descriptive questions and answers, suitable for applications such as learning assistants or visual content accessibility systems.

The data flow is described in detail in Figure 2 and consists of the following key steps:

+ Photo data + caption → the question (T5TP3).
+ Create a descriptive VQA dataset (image, question, answer).
+ Fine-tune BLIP-VQA with prompt engineering.

+ Input: image (ViT), question (tokenized).
+ Output: descriptive answer.

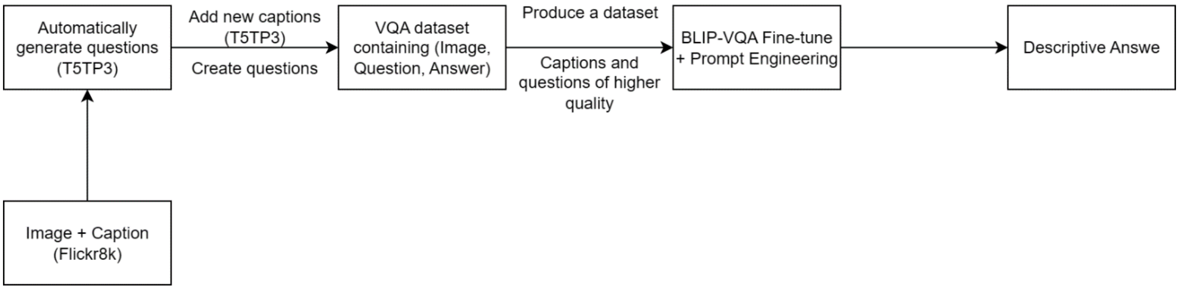
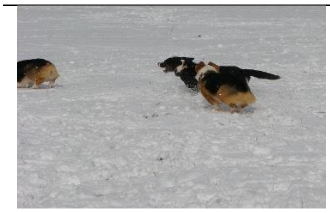

Automatically generate questions (T5TP3)
Add new captions (T5TP3)
Create questions
VQA dataset containing (Image, Question, Answer)
Produce a dataset
Captions and questions of higher quality
BLIP-VQA Fine-tune + Prompt Engineering
Descriptive Answe
Image + Caption (Flickr8k)

Figure 2. Data flow description

Table 1. Some examples of comparisons between the pre- and post-fine-tune models.

Image	Question	Reply to original BLIP	Reply after fine-tune
	How many dogs are playing around in the snow?	"Three"	"Three dogs are playing around in the snow."
	What happens when a child jumps into a pool?	"Jumping"	"A child is jumping into a pool while a man is watching the child"

Prompting Techniques

During the training and assessment process, we followed prompt methods presented in [29-30] 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. Table 1 shows some examples of comparisons between the pre- and post-fine-tune models.

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.

3.3. Model Fine Tuning

In this section, we describe the process of answering questions based on image content using the fine-tuned BLIP model in combination with Prompt Engineering techniques. The system employs BLIP-VQA as the core question–answering module, which enables efficient fusion of visual and textual information. We selected the BLIP framework for its compact yet powerful end-to-end architecture, capable of performing image–text alignment, feature extraction, and answer generation within a unified model. When combined with Prompt Engineering, BLIP becomes more adaptive to various question types and linguistic patterns, allowing the system to optimize query interpretation and produce descriptive, contextually relevant, and informative answers that reflect a deeper understanding of visual scenes.

The fine-tuned BLIP model leverages hierarchical co-attention and the Stacked Attention Network (SAN) to strengthen the interaction between image regions and textual representations. These mechanisms enhance the model's ability to capture fine-grained spatial dependencies and semantic relationships across modalities. Although traditional attention-based approaches often depend heavily on localized image features and struggle with multi-step reasoning or contextual inference, our integration of BLIP with Prompt Engineering mitigates these limitations. By dynamically refining prompts during the inference process, the system can better focus on relevant visual regions and linguistic cues, improving its capacity for compositional reasoning and detailed description generation. Consequently, this approach provides a balanced trade-off between computational efficiency, model interpretability, and semantic depth, enabling the VQA system to deliver answers that are both accurate and naturally articulated.

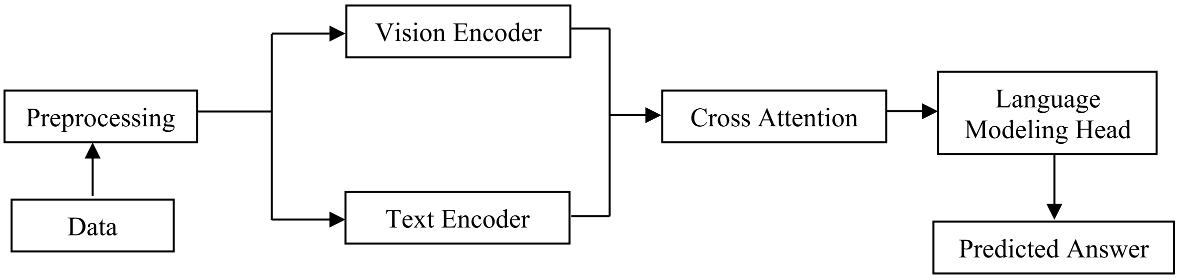


Figure 3. Detailed description of the fine-tuning process of BLIP

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 decribes detailed description of the fine-tuning process of BLIP. This processing includes following steps:

Forward pass

When using HuggingFace's Trainer to fine-tune BLIP for the VQA problem, the internal forward pass process takes place as follows:

- Data input: Images are read from the DataLoader and converted into pixel_values tensors according to BLIP's standards, and the question is tokenized (inputids, attentionmasks) using BLIP's tokenizer.

- Characteristic extraction from images: Vision Encoder (usually Vision Transformer – ViT receives image tensors and transforms them into a series of image embeddings, which store high-level visual information (objects, colors, etc.).

- Characteristic extraction from the question: The Text Encoder (usually based on Transformer) receives the question's token string and generates text embeddings.

- Combining image and text information: BLIP uses Cross Attention Layers to allow text embedding "attention" to the relevant areas of the image, and vice versa, the result is a multi-modal representation that is full of context and relevant details.

- Predictive Answer Generating: Multi-modal representation is included in the Language Modeling Head (Linear layer + Softmax) to generate a probability for each answer token. The model picks out the final predicted answer.

Loss calculation

The trainer calls the compute_loss function of BlipForQuestionAnswering, where:

- Compare predictions and ground truths at the token level.

- Use Cross-Entropy Loss to calculate the difference between the predicted probability distribution and the true label.

- Mask tokens padding so that losses are only calculated on valid tokens.

This loss represents the overall deviation between the predicted answers and the standard answers on the current batch.

Backpropagation

- Backward pass: Based on the gradient of the loss, the Trainer asks PyTorch to calculate the derivative for all parameters in the Vision Encoder, Text Encoder, and Cross Attention Layers.

- Weight Update: Use the AdamW optimizer (Adam with weight decay) to adjust the weights in the direction of reducing losses.

- Important metaparameters: Learning rate is used to control the update rate, Weight decay aims to reduce overfitting.

- Scheduler: Trainers often use Linear Warmup & Decay, which helps the learning rate gradually increase at the beginning of the training to avoid instability, and then gradually decrease at the end so that the model is better converged.

3.4. Evaluation

We use evaluation indicators that are suitable for the problem of generating automatic answers for VQA: BLEU [31] and ROUGE-L [32].

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 Eq (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: Calculate the n-gram match ratio and apply clipping to avoid repeating the word fraud.

- 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.

$$R_{LCS} = \frac{LCS(hypo,ref)}{m} \quad (2)$$

$$P_{LCS} = \frac{LCS(hypo,ref)}{n} \quad (3)$$

ROUGE-L is calculated as follows:

$$ROUGE\text{-}L = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}} \quad (4)$$

where:

LCS(hypo,ref): the length of the longest consecutive substring between the birth and reference sentences.

m: the length of the reference sentence (*ref*).

n : the length of the sentence generated (*hypo*).

β : is a weighted adjustment parameter between Recall and Precision.

In addition, we use the following measurements to evaluate the question generated:

Mean Question Similarity: The average of similarity (in terms of cosine similarity or a similar index) between questions generated in the same data set. Evaluate the diversity and relevance between questions.

Mean Question–Caption Similarity: Average the similarity between the generated question and the corresponding original caption. Evaluate the relevance and accuracy of the question compared to the content of the caption.

Unique Question Ratio: The percentage of questions generated that are unique (not repeated) out of the total number of questions. Evaluate the diversity and creativity of the question generation system.

4. RESULTS AND DISCUSSION

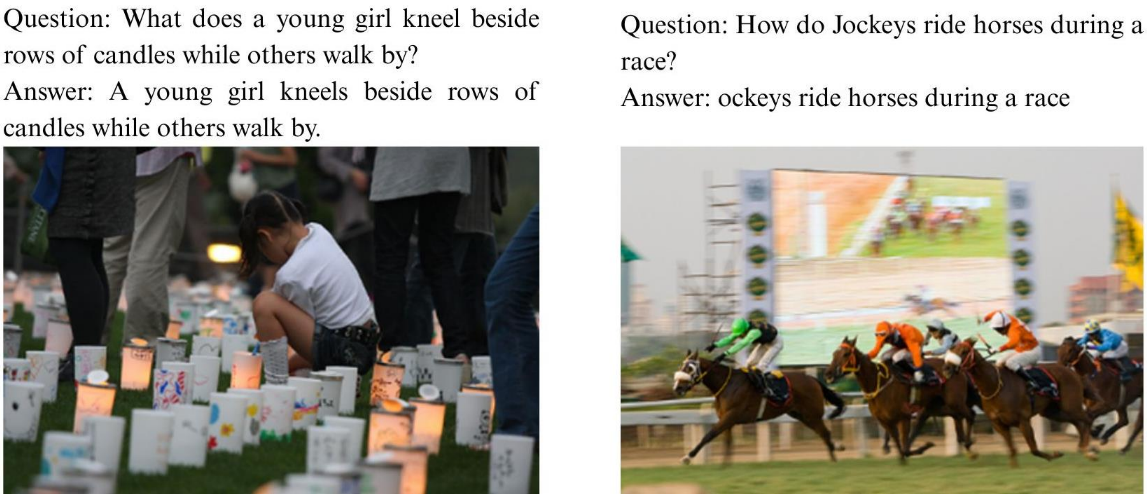
4.1 Dataset

The dataset used in the experiment was Flickr8k [33], which consisted of 8000 images depicting situations in everyday life. Each photo has several short captions in English. Some sample data is shown in Figure 4. 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).
+ Each template includes: image, question, answer, and is saved to a file that makes up the fine-tuned Flickr8k dataset.

Table 2. Dataset statistics

Dataset	Number
Train set	5663
Test set	2428



Question: What does a young girl kneel beside rows of candles while others walk by?
Answer: A young girl kneels beside rows of candles while others walk by.

Question: How do Jockeys ride horses during a race?
Answer: ockeys ride horses during a race

Figure 4. Some sample data

4.2. 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.

In this study, we fine-tune the BLIP-VQA (Salesforce/blip-vqa-base) model on the Flickr8k dataset reprocessed as Visual Question Answering. 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 [34], InstructBLIP [35].

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-Finetune (ours)	**0.32**	**0.27**	**0.23**	**0.19**	**0.52**

After fine-tuning the BLIP-VQA 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 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.

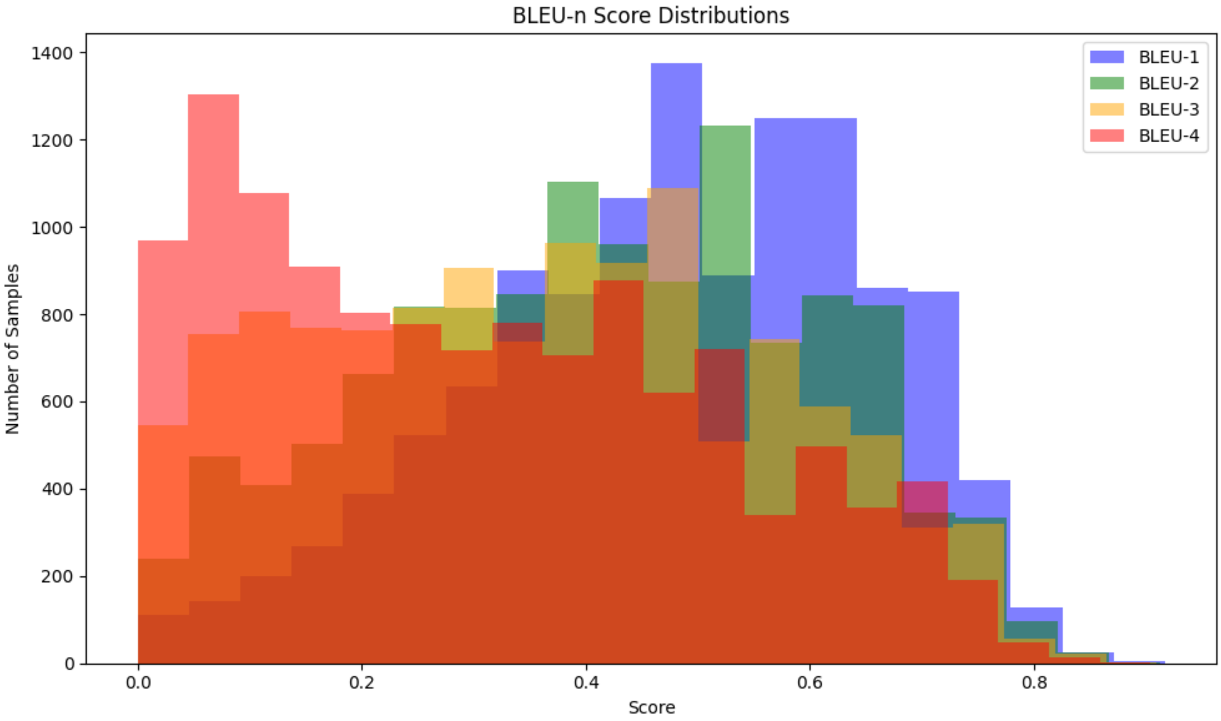


Figure 5. BLEU-1 to BLEU-4 score distribution chart on the test dataset

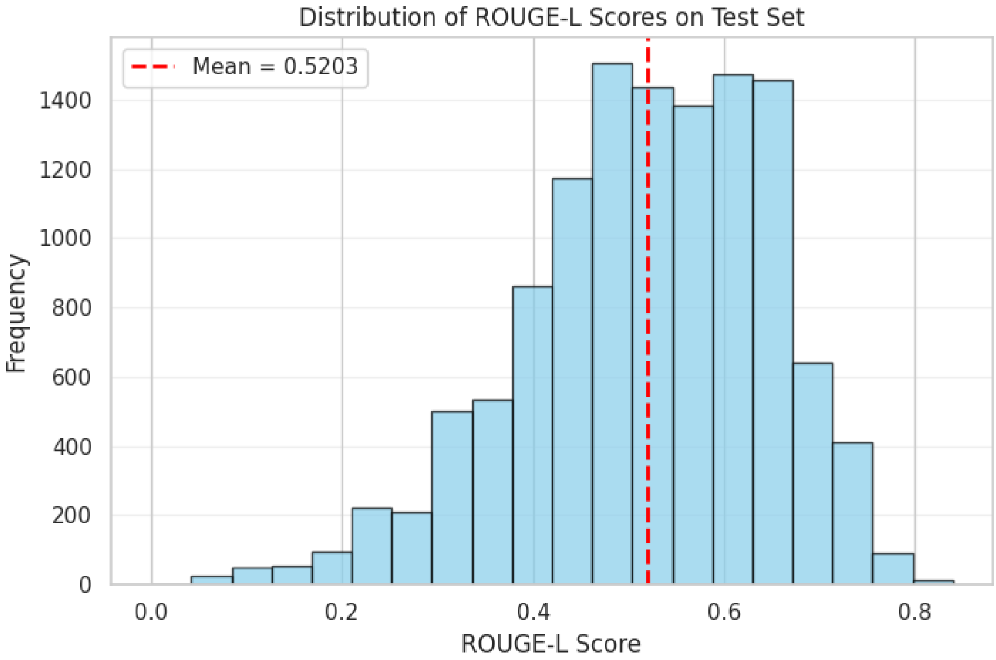


Figure 6. ROUGE-L distribution chart on test set

Figure 5 shows the BLEU-n score distribution (with n = 1 to 4) of the BLIP model after fine-tune on the VQA dataset described. It can be seen that BLEU-1 has the highest value, concentrated mainly in the range of 0.5–0.7, indicating that the model is capable of generating single words that are accurate compared to the reference sentence. Meanwhile, BLEU-2 and BLEU-3 declined, reflecting a moderate level of short-contextual relevance (bigram, trigram). BLEU-4 is strikingly distributed towards the low point, mostly in the range of 0.1–0.3, suggesting that the model still has difficulty producing long sequences that match the ground-truth exactly. This shows that the model's answers often match only a small part of the original sentence, and are susceptible to bias when dealing with long or complex patterns both grammatically and semantively. However, this is also a common phenomenon in automated text generation systems, where the model often prioritizes generating semantic answers rather than absolute matches to the reference sentence.

Question: How many dogs are on a beach looking at a red ball in someones hand?
Answer: Three dogs on a beach looking at a red ball in someones hand
Answer prediction: Three dogs on a beach looking at a red ball in someone ' s hand.

Question: Who do park their red motorcycles near a church?
Answer: Two men park their red motorcycles near a church
Answer prediction: A man in a red uniform and a man in a black uniform park their red motorcycles near a church

Question: Who is resting after a game of fetch?
Answer: The dog is resting after a game of fetch.
Answer prediction: a black and brown dog is resting after playing fetch.

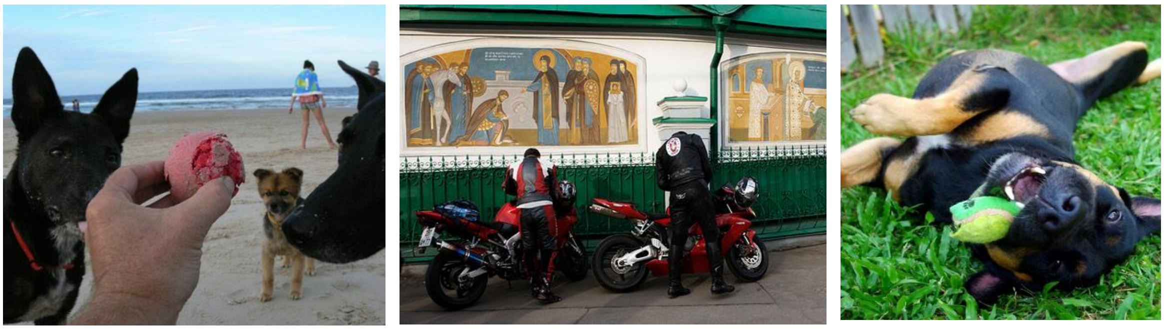


Figure 7. Results of some samples from the test set.

Question: How many people are walking toward the mountains?
Answer: A group of 5 people are walking toward the mountains
Answer prediction: Four people are walking toward the mountains.

Question: Several people dressed in jerseys roll an old cannon during a competition?
Answer: Several people dressed in like jerseys roll an old cannon during a competition.
Answer prediction: several people wearing different colored shirts are gathered around a tire swing.

Question: What is the name of Oklahoma University's Sooners football team?
Answer: Oklahoma University 's Sooners football team playing against a rival.
Answer prediction: a football player in a red and white uniform is running down the field.

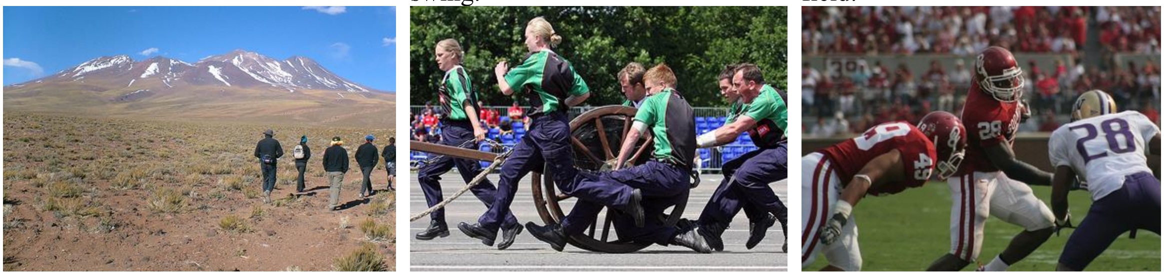


Figure 8: Some patterns are mispredicted

Figure 6 reflects the distribution of ROUGE-L points on the test set with a near-standard pattern but with a slight deviation to the left. The majority of the sample is concentrated in the range of 0.4 to 0.65, suggesting that the model often produces answers that are quite similar to the original answers. An average value of around 0.52 (indicated by a red line) indicates that the overall efficiency of the model is at an acceptable level. However, the distribution was still widespread: some samples scored very low (below 0.2), while others approached 0.8. This reflects inconsistent performance - the model can handle some types of questions well but struggles with others, perhaps due to semantic differences or contextual complexity. The appearance of high scores is a positive signal, confirming the ability to generate text that is close to the original meaning, but there is still room for improvement to improve consistency and generalization across the entire dataset.

Figure 7 shows the results of some samples from the Test set, and Figure 8 shows some samples that were incorrectly predicted. False prediction errors such as Figure 8 in the above samples indicate that the BLIP model has some obvious limitations. First of all, the model cannot answer questions that require background knowledge other than images, such as the name of the "Oklahoma University" football team, because there is no knowledge retrieval or text recognition (OCR) mechanism. Second, the model does not accurately count the number of people or correctly identify complex emotions such as "excited", due to a lack of ability to quantify and understand the context of expression. Finally, the model is prone to confusing objects or actions in the image – for example, mistaking "cannon" for "tire swing", indicating weak ability to identify specialized objects and infer collective action. These limitations are mainly due to the fact that the model relies solely on images and input questions without the support of external sources of information or specialized modules such as counting, emotion recognition, or background knowledge.

5. CONCLUSION

This study proposed a method to improve interoperability and automatic answer generation in the Visual Question Answering system by combining the BLIP-VQA model with an automated question generation pipeline. We built a VQA dataset from the Flickr8k set, using the T5TP3 model to generate questions from the original caption, and fine-tuning BLIP-VQA to generate descriptive answers.

The experimental results showed that the proposed method achieved higher performance than previous models such as BLIP-2, InstructBLIP with BLEU and ROUGE-L scores with significant differences. Analysis of the BLEU-1 to BLEU-4 scores also indicates that the model is capable of reproducing single words accurately, but struggles with longer sequences of words. Similarly, the ROUGE-L score shows that the model can produce an answer with a fair degree of moderation similarity to the standard answer, but there are still fluctuations between data samples.

Despite many improvements, the BLIP-VQA model still has some limitations. Specifically, the system has not been able to answer questions that require background knowledge in addition to images, has difficulty recognizing the number of objects and complex emotions, and is confused between objects or actions. These limitations mainly stem from the fact that the model relies solely on image data and input questions without the support of additional modules such as OCR, emotion recognition, or background knowledge.

This study demonstrated that the fine-tune BLIP-VQA method has the potential to significantly improve the ability to automatically answer image questions. The findings from the experiment could contribute to the development of smarter VQA systems, which will enhance the experience of human-computer interaction in computer vision and natural language processing applications.

In the future, in order to further improve the quality of VQA, we will look at improvements such as tuning the question generation algorithms to increase the variety and accuracy of the input questions, expanding the training dataset by combining various image and text sources to increase the generalization of the model and apply the Use Prompt Engineering to improve questions and answers for VQA.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

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	✓	✓			✓	✓		✓	✓	✓		✓	✓	✓

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 [33].

REFERENCES

[1] Bui Thanh Hung, Ho Vo Hoang Duy. ExVQA: A novel Stacked Attention Networks with Extended Long Short-Term Memory Model for Visual Question Answering. Computers and Electrical Engineering. Volume 126, 3 August 2025, 110439, 1-20. 2025. ISSN: 0045-7906. https://doi.org/10.1016/j.compeleceng.2025.110439

[2] Bui Thanh Hung. Vietnamese Question Classification based on Deep Learning for Educational Support System. The 19th International Symposium on Communications and Information Technologies, ISCIT 317-321, 2019. https://doi.org/10.1109/ISCIT.2019.8905237

[3] S. Yang, Z. Li, Y. Xu, J. Tang, and G. Huang, "MAGIC-VQA: Multi-hop and Grounded Inference for Commonsense Visual Question Answering," arXiv preprint arXiv:2502.06745, 2025. https://doi.org/10.48550/arXiv.2503.18491

[4] S. Ali, M. U. Khan, and F. S. Khan, "A Journey Through the Evolution of Visual Question Answering: From Rule-Based to Multimodal Transformers," *ACM Comput. Surv*., vol. 57, no. 1, pp. 1–37, 2025. https://doi.org/10.48550/arXiv.2501.07109

[5] M. Tan, L. Liu, Y. Chen, and P. Wang, "Visual Question Answering: A Survey of Methods, Datasets, and Evaluation Metrics," ACM Trans. Intell. Syst. Technol., vol. 16, no. 2, pp. 1–41, 2025. https://dl.acm.org/doi/10.1145/3728635

[6] W. Chen, J. Zhang, and Q. Yu, "A Comprehensive Review on Multi-Modal Fusion Techniques in Visual Question Answering," *Pattern Recognit. Lett*., vol. 167, pp. 35–45, 2023. https://doi.org/10.48550/arXiv.2408.01319

[7] Chowdhury, S., & Soni, B. R-VQA: A robust visual question answering model. *Knowledge-Based Systems*, 309, 112827. 2025. https://doi.org/10.1016/j.knosys.2024.112827

[8] Y. Gao, H. Li, and S. Li, "Visual Question Answering: From Early Developments to Large Multimodal Models," IEEE Trans. Pattern Anal. Mach. Intell., 2025. https://doi.org/10.48550/arXiv.2501.03939

[9] Yamada, M., D'Amario, V., Takemoto, K., Boix, X., & Sasaki, T. Transformer module networks for systematic generalization in visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10096-10105. 2024. https://doi.org/10.48550/arXiv.2201.11316

[10] Z. Wang, Y. Huang, H. Wang, and Y. Li, "MIRTT: Multimodal Interaction via Recurrent Trilinear Transformers for Visual Question Answering," Pattern Recognit., vol. 119, p. 108047, 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.196

[11] Qian, Y., Hu, Y., Wang, R., Feng, F., & Wang, X. Question-driven graph fusion network for visual question answering. In 2022 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE. 2022. https://doi.org/10.48550/arXiv.2204.00975

[12] Kabir, R., Haque, N., & Islam, M. S. A comprehensive survey on visual question answering datasets and algorithms. arXiv preprint arXiv:2411.11150. 2024. https://doi.org/10.48550/arXiv.2411.11150

[13] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "IQA: Visual Question Answering in Interactive Environments," arXiv preprint arXiv:1712.03316, 2017. https://doi.org/10.48550/arXiv.1712.03316

[14] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 2425–2433. 2015. https://doi.org/10.48550/arXiv.1505.00468

[15] de Faria, A. C. A. M., Bastos, F. D. C., da Silva, J. V. N. A., Fabris, V. L., Uchoa, V. D. S., Neto, D. G. D. A., & Santos, C. F. G. D. Visual question answering: A survey on techniques and common trends in recent literature. arXiv preprint arXiv:2305.11033., 2023. https://doi.org/10.48550/arXiv.2305.11033

[16] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Comput. Vis. Image Underst*., vol. 163, pp. 3–20, 2017. https://doi.org/10.1016/j.cviu.2017.06.005

[17] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. "Visual question answering: A survey of methods and datasets". *Computer Vision and Image Understanding*, 163, 21-40. 2017. https://doi.org/10.48550/arXiv.1607.05910

[18] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances in Neural Information Processing Systems (NeurIPS), vol. 29, 2016. https://doi.org/10.48550/arXiv.1606.00061

[19] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 21–29. https://doi.org/10.48550/arXiv.1511.02274

[20] Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., & Zhou, M. Visual question generation as dual task of visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6116-6124). 2018. https://doi.org/10.48550/arXiv.1709.07192

[21] Bui Thanh Hung. Content based Image Retrieval using Multi-Deep Learning Models. Next Generation of Internet of Things. Lecture Notes in Networks and Systems-LNNS. Springer, Singapore. ISSN: 2367-3370, vol 445. pp. 347-357, 2022. https://doi.org/10.1007/978-981-19-1412-6_29

[22] Bui Thanh Hung. Link Prediction in Paper Citation Network based on Deep Graph Convolutional Neural Network. Computer Network, Big Data and IoT. Lecture Notes on Data Engineering and Communications Technologies- LNDECT. Springer, Singapore. ISSN: 2367-4512, vol 117, pp. 897–907, 2022. https://doi.org/10.1007/978-981-19-0898-9_67

[23] Bui Thanh Hung, Vo Quoc Huy. EMTFIC: enhanced fashion image captioning via multi-transformer architecture with contrastive and bidirectional encodings. The Visual Computer Journal. 1-15, 2025. ISSN: 0178-2789. https://doi.org/10.1007/s00371-025-04072-8

[24] Bui Thanh Hung, Nguyen Van Phuc Nhan, Nguyen Tien Sy. DCARES: deep convolutional neural network with neural-based optimization for image-based product recommender system. Multimedia Tools and Applications Journal, 2025. ISSN: 1573-7721. https://link.springer.com/article/10.1007/s11042-025-20655-y

[25] K. Q. Tran, A. T. Nguyen, A. T. H. Le, and K. Van Nguyen, "ViVQA: Vietnamese visual question answering," in Proc. 35th Pacific Asia Conf. Language, Inf. Comput., 2021, pp. 683–691. https://aclanthology.org/2021.paclic-1.72/

[26] N. H. Nguyen, D. T. D. Vo, K. V. Nguyen, and N. L.-T. Nguyen, "OpenViVQA: Task, Dataset, and Multimodal Fusion Models for Visual Question Answering in Vietnamese," arXiv preprint arXiv:2305.04183, 2023. https://doi.org/10.1016/j.inffus.2023.101868

[27] J. Li, D. Li, C. Xiong, and S. C. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," arXiv preprint arXiv:2201.12086, 2022. https://doi.org/10.48550/arXiv.2201.12086

[28] C. Zhang, H. Zhang, and J. Wang, "Downstream Transformer Generation of Question-Answer Pairs with Preprocessing and Postprocessing Pipelines," arXiv preprint arXiv:2205.07387, 2022. https://doi.org/10.48550/arXiv.2205.07387

[29] Y. Yang, Y. Wang, D. Wang, J. Gao, and J. Liu, "PromptBench: Towards evaluating the robustness of prompts for vision and language models," arXiv preprint arXiv:2201.12086, 2022. https://doi.org/10.48550/arXiv.2306.04528

[30] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927. 2024. https://doi.org/10.48550/arXiv.2402.07927

[31] Kim, B. S., Kim, J., Lee, D., & Jang, B. Visual question answering: A survey of methods, datasets, evaluation, and challenges. ACM Computing Surveys, 57(10), 1-35. 2025. https://doi.org/10.1145/3728635

[32] Ma, J., Wang, P., Kong, D., Wang, Z., Liu, J., Pei, H., & Zhao, J. Robust visual question answering: Datasets, methods, and future challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8), 5575-5594. 2024. https://doi.org/10.48550/arXiv.2307.11471

[33] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., & Lazebnik, S. Improving image-sentence embeddings using large weakly annotated photo collections. In European conference on computer vision (pp. 529-545). Cham: Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-10593-2_35

[34] H. Li, P. Zhang, J. Li, X. Qi, L. Wang, and L. Zhang, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 12243–12254. 2023. https://doi.org/10.48550/arXiv.2301.12597

[35] W. Dai, Z. Yang, Z. Zhang, K. Lin, J. Han, and X. Lin, "InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning," arXiv preprint arXiv:2305.06500, 2023. https://doi.org/10.48550/arXiv.2305.06500

BIOGRAPHIES OF AUTHORS

	Nguyen Ha Manh Khang He 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 He 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 He 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 He received his M.S. degree and Ph.D. degree from Japan Advanced Institute of Science and Technology, Japan (JAIST) in 2010 and in 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 Data Science Laboratory, Data Science Department, 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.