A MULTIMODAL TRANSFORMER-BASED ARCHITECTURE FOR VIETNAMESE IMAGE CAPTIONING FOCUSING TRAFFIC SCENES

Van Dung Hoang[1], Duc Nhan Bui[1,*], Thien Nhan Huynh[1], Trieu Phi Dai[1]

DOI: https://doi.org/10.57001/huih5804.2025.414

ABSTRACT

This study presents a multimodal image captioning framework tailored for the Vietnamese traffic domain. The architecture combines Vision Transformer for visual feature extraction, BARTpho for Vietnamese language encoding, and GPT-2 for caption generation. A traffic-related dataset is constructed via web scraping and automatically labeled using a large language model. Image augmentation is employed to improve object diversity and generalization. Experimental results show that the model produces context-aware, coherent captions and achieves high evaluation scores, highlighting its potential for smart transportation, urban surveillance, and assistive technologies.

Keywords: *Vietnamese Image Captioning, Multimodal Learning, Vision-Language Transformer, Deep Neural Network, Vietnamese Dataset.*

[1]Ho Chi Minh City University of Technology and Education, Vietnam
[*]Emai: nhanbui15122003@gmail.com

Received: 22/8/2025
Revised: 10/11/2025
Accepted: 28/11/2025

1. INTRODUCTION

Image captioning has been a very popular topic in the computer vision field recently. Image captioning aims at describing the content of an image using descriptive text. It is an interesting and challenging problem because of its diverse applications in real life and also because of the difficulty in making computer vision systems understand the content of an image. Moreover, the creation of image captioning systems has tremendous potential to help visually impaired individuals "see" the world through descriptive text.

Computer Vision (CV) and Natural Language Processing (NLP) used to be two separate research domains. However, with the booming growth of multimedia data from mobile devices, IoT systems, and social media, researchers now tend to merge techniques from the two fields for improved understanding of multimodal content and more efficient extraction of useful information from visual data. Actually, automatic image annotation a key task in image retrieval and computer vision is meant to attach meaningful words or phrases to represent images. To this end, artificial intelligence approaches, typically with the help of pre-trained models, are used to discover mappings from low-level visual features to high-level semantic representations for generating meaningful captions for a given image.

Image captioning has been a flagship task at the intersection of computer vision and natural language processing, and top-down and bottom-up have been two prominent paradigms. Whereas the top-down approach takes an image, extracts the global visual features, and translates them into natural language, the bottom-up approach begins with identifying and describing salient regions in the image, which are integrated into consistent textual descriptions. Both approaches typically rely on advanced language models to supply fluency, coherence, and contextual relevance of the generated captions.

In this paper, we propose a novel image captioning model using the top-down method, enhanced with an Attention Mechanism within the encoder-decoder paradigm. With this, the model is capable of selectively focusing on the most informative visual cues when creating explanatory captions. Even though current models have achieved state-of-the-art results for English, there have been limited studies on Vietnamese image captioning in traffic scenes. To bridge this gap, we test the proposed model on a Vietnamese image captioning dataset of traffic scenes. Our work is aimed not only at advancing image captioning research but also at

supporting real-world applications. We are targeting the development of assistive technology for the blind, with the particular aim of enhancing their situational awareness and traffic safety by describing images in Vietnamese automatically.

The remainder of this paper is structured as follows: the remainder of Section 2 discusses related works from traditional to state-of-art ones, specifying model's output goals and coverage; Section 3 explains the suggested approach to build Vietnamese image captioning dataset and multimodal framework underpinning the research; Section 4 presents the final outcome and 5 for conclusion.

2. MATERIALS AND METHODS

There are some available image datasets of traffic scenes that are typically built for general-purpose usage such as object detection or scene segmentation and tend to be primarily focused on well-structured data captured in ideal conditions. Within the specific case of traffic scenes, open-source captioning corpora remain absent and those available are typically limited in contextual information, typically supplying short object-based descriptions rather than semantically rich stories that include real-world interaction and environmental signals. Besides, there is also the absence of high-quality Vietnamese-language traffic image captioning datasets while significant progress has been made in creating image captioning datasets, most of the existing resources are targeted towards general domains such as indoor scenes, everyday objects, or social events with a strong English-language caption preference. This language disparity represents a major stumbling block in developing AI systems that can communicate with non-English-speaking users, particularly in critical applications such as accessible technology for the visually impaired, where accurate and context-aware scene perception is critical.

To address these limitations, we propose creating a domain-specific, high-quality image captioning dataset called the Traffic Pictures Captioning dataset [14] also known as TPC37k, which depicts the complexity and diversity of real-world traffic situations. This dataset includes a wide range of traffic-related visual content, such as object categories (vehicles, pedestrians, traffic signs, lights, sidewalks, roads), lighting conditions (daytime, nighttime, dusk, weather variations), perspectives (wearable cameras, traffic surveillance), and traffic scenarios (intersections, one-way streets, residential areas, and so on).

2.1. Building Dataset

The TPC37k dataset [14], as depicted in Figure 1, was constructed through a systematic pipeline (refer to Figure 1). Traffic-related images were initially collected from web pages using SerpApi with traffic-specific keywords for topic diversity and relevance. Data preprocessing was subsequently performed to verify image URLs, remove duplicates (approximately 10%), and remove low-resolution images, resulting in a clean dataset of 9,300 images. Automatically generated initial captions came from the Gemini 2.0 Flash API and were subsequently edited manually (80% of captions) to ensure linguistic accuracy and contextual relevance. Finally, various augmentation techniques (e.g., pixel shift, spatial transformation, adding noise) expanded the dataset to 37,056 images, adding data richness and enabling strong model training.

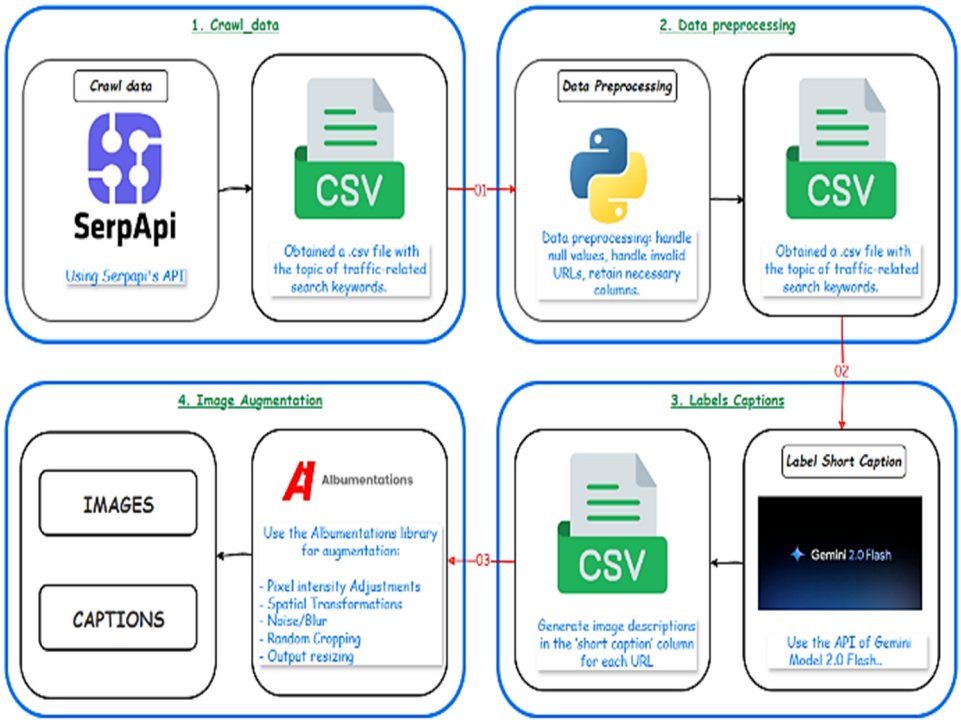

Figure 1. The overall pipeline for building TPC37k dataset

Table 1. Token and word count in TPC37K

Type	Total	Unique
Token	1,160,004	1,731
Word	1,267,648	2,356

Table 2. The plits in TPC37K

Split	Images	Captions	Images/ Caption	Max length	Avg. length
Train	29,644	7266	4	71	31,34
Val	3,706	922	4	64	31,12
Test	3,706	920	4	58	31,16

2.1.1. Data Augmentation

In order to increase the diversity and scalability of the dataset, we apply various image enhancement techniques using the Albumentations library:

• Pixel-level transformation: Our approach is that each input image has a 50% chance of being altered in brightness and contrast (within ±20%) or being colored with variations in brightness, contrast, saturation (±20%) and hue (±0.1).

• Spatial transformations: In order to simulate different viewing angles and small geometric variations, translation-scale-rotation operations (translation and scale within ±10%, rotation limit 15°) or affine transformations (scale from 0.9 to 1.1, translation within ±10% and rotation from -10° to 10°) were applied with a probability of 50%.

• Noise and blurring: In order to reproduce common defects in real images, Gaussian noise (with variance from 10 to 30) or Gaussian blurring (with kernel size from 3 to 5) were randomly applied with a probability of 30%.

• Random cropping and resizing: Randomly resized image cropping is performed with a ratio range of 0.8 to 1.0 and an aspect ratio of 0.9 to 1.1 to simulate the occlusion of the image when it is actually captured. All images are then resized down to a fixed resolution of 512x512 pixels to ensure consistency across the entire dataset.

2.1.2. Creating caption labels

To create caption labels for images, we use the Gemini 2.0 Flash model through the API provided by Google AI Studio. The captions are created in Vietnamese (through a properly adjusted prompt), ensuring the necessary constraints (maximum 40 words per caption), describing the main issues and generalizing the image content accurately. The Prompt technique is continuously adjusted to refine the labeling behavior so that the captions are consistent with human descriptions and suitable for guiding the visually impaired.

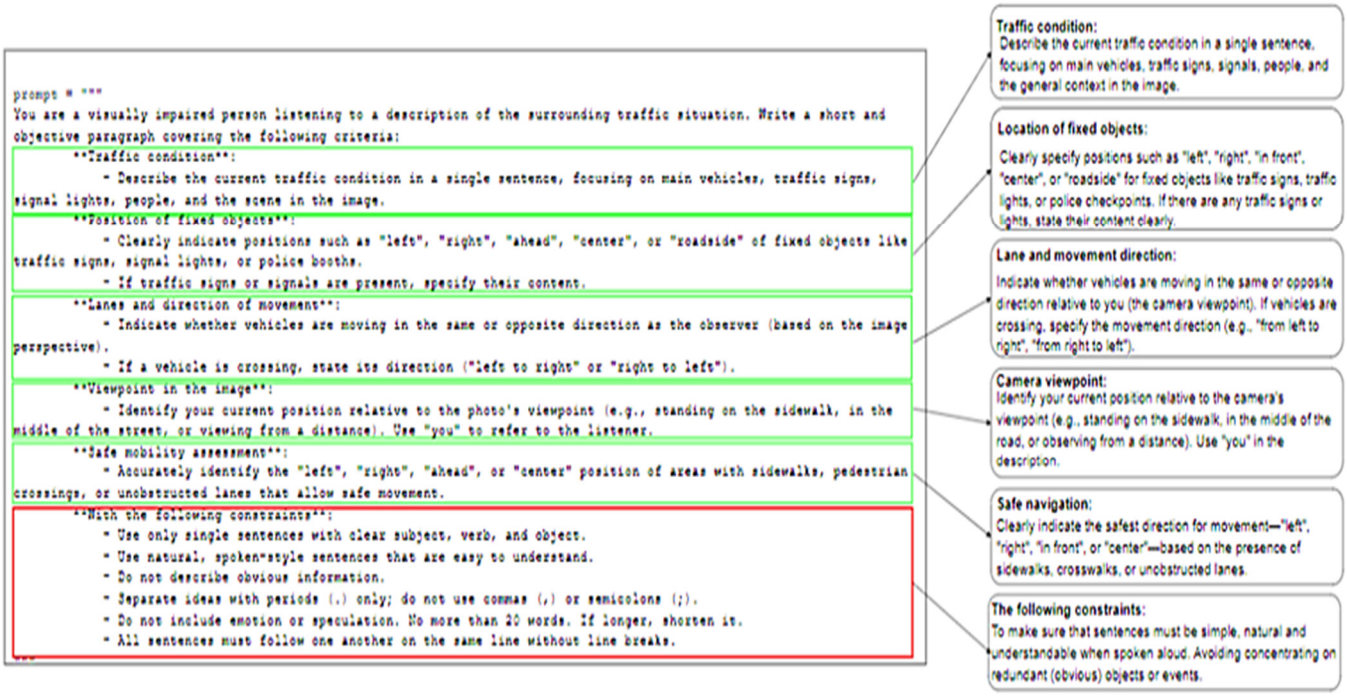

Figure 2. Prompt template for caption label generating

2.1.3. Data Preprocessing: In order to ensure the quality and consistency of input data for image captioning modeling, we designed a systematic data processing pipeline (see Figure 3) that involves multiple interdependent stages, as depicted in Figure 1. The pipeline is designed to process both image and caption

data in parallel and make them clean, standardized, and semantically consistent before being used for model training.

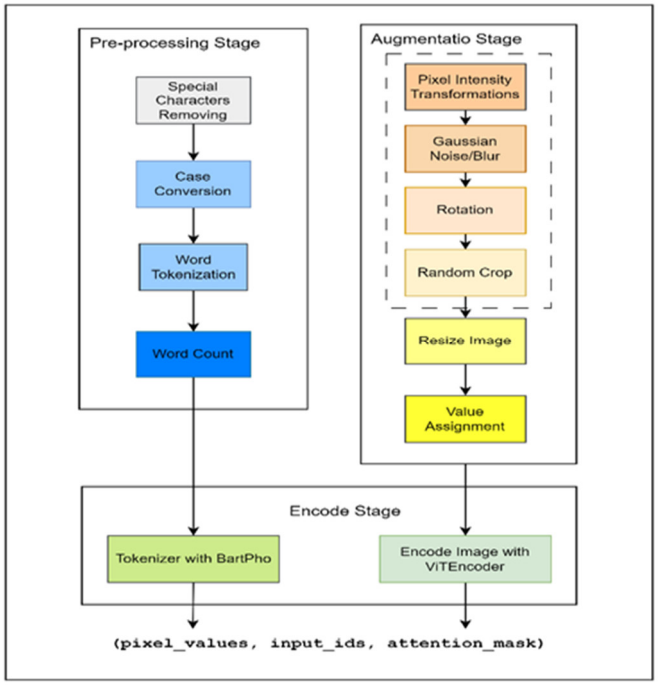

Figure 3. Data preprocessing pipeline for caption images to training structure

It begins with image crawling for raw image-caption pairs acquisition, which are sanitized to remove noise and formatting anomalies. The image-augmentation operation of resizing, cropping, and blurring is used to increase model robustness by introducing visual variation from the same source material.

These are preprocessed using the ViT processor (google/vit-base-patch16-224), which rescales input to the size 224×224 and splits them up into 16×16 patches. Pixel intensities are standardized using ImageNet statistics, returning tensors of the shape $(1, 3, 224, 224)$ that conform to the input format of the ViT encoder. Besides, captions are tokenized with the BARTpho word-level tokenizer (vinai/bartpho-word), which includes special tokens that are needed and word-to-vocabulary-ID mappings. Sequences are padded or truncated to 315 tokens at most and become tensors of shape *(batch_size, max_length)* for GPT-2.

Both image and caption tensors are aligned inside the combiner module, which allows proper matching. They are thereafter passed to the encoding stage where abstractions are extracted from features to serve as inputs to the captioning model. This modularity allows for clean and uniform multimodal preparation for training.

2.2. Proposed Architecture for Image Captioning PVG

Image captioning involves two key components: extracting semantic information from visual content and generating grammatically correct natural language descriptions. In this work, we propose an image captioning framework based on the widely adopted Encoder-Decoder architecture. As can be seen in Figure 4, our approach leverages pretrained models, including Vision Transformer (ViT) [11] as the image encoder and GPT-2 as the language decoder both of which are Transformer-based architectures. Because our focus is on the Vietnamese language, we utilize the BARTpho tokenizer [15], a word-level tokenizer specifically pre-trained on Vietnamese high-scale datasets. Even though originally designed for BART-based models, BARTpho tokenizer can be used with Transformer-based decoders such as GPT-2 and supports effective tokenization tailored to Vietnamese linguistic patterns. The resulting model consists of two main modules: a visual encoder and a caption decoder. We refer to our proposed architecture as BartPho-ViT-GPT2, and for brevity, we will denote it as PVG in the remainder of this paper.

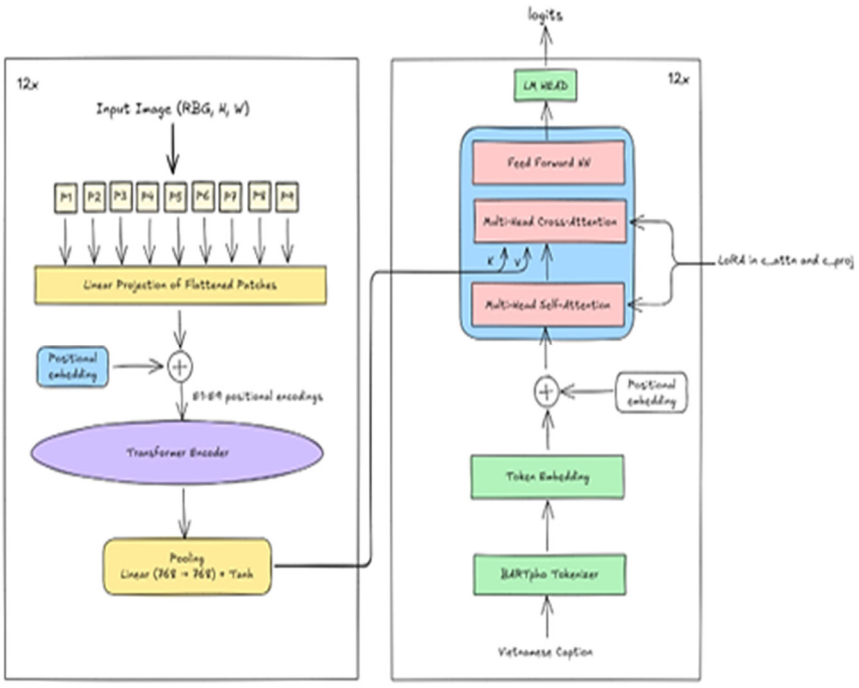

Figure 4. Multimodal architecture following Encoder-Decoder of PVG

The ViT and GPT-2 pretrained models, with the assistance of the BARTpho tokenizer, constitutes an interesting multimodal approach to image captioning, particularly for Vietnamese text generation.

Unlike conventional CNN-RNN architecture, the Vision Transformer (ViT) encoder captures global spatial relationships across the entire image via self-attention, instead of relying on local receptive fields. This enables a deeper understanding of the visual context, which is crucial for image captioning. Moreover, ViT's modular design, which avoids handcrafted components like anchor boxes, facilitates seamless integration with Transformer-based decoders. The visual input I is encoded into a latent representation $X = \{x_1, x_2, \ldots, x_k\}$ where each x_i represents the feature vector of an image patch. The resulting visual embedding $V \in R^{n \times d}$ with n being the number of patches and d the hidden dimension, is used as input to the language model. The model then predicts each caption token y_n based on prior tokens and the visual context, as shown in equation (2).

$$X = \text{Encoder}(I) \quad (1)$$

$$P(y_1, y_2, \ldots, y_n | X) = \prod_{t=1}^{T} P(y_i | y_1, y_2, \ldots, y_n, X) \quad (2)$$

ViT easily integrates with Transformer-based decoders due to their shared architectural principles. The output from the [CLS] token is linearly projected and passed through a Tanh activation function to generate the input embedding for the decoder.

$$\widetilde{X} = \tanh(W \cdot h_{[CLS]} + b) \quad (3)$$

Where $W \in R^{d \times d}$ and $b \in R^{d}$. This transformation outputs the final image embedding used as input for the decoder.

We use the BARTpho-word tokenizer to process Vietnamese text. Operating at the word level and pre-trained on Vietnamese data, BARTpho helps reduce word segmentation errors, resulting in more linguistically appropriate word representations.

After encoding, the $input\_ids$ and $attention\_mask$ are fed into the $self - attention$ layer to model intra-caption dependencies. The output of $self - attention$ serves as the $query\ (Q)$ in the $cross - attention$ mechanism. The visual features ($pixel\_values$) extracted by the encoder (ViT) are used as the key (K) and value (V). The cross-attention layer integrates visual information with the context of the caption. These can be formulated as Equations (4) and (5):

$$Q = W^{Q} \times X_{text};\ K = W^{K} \times X_{image};$$
$$V = W^{V} \times X_{image} \quad (4)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \quad (5)$$

The architecture is constructed by integrating pre-trained models, taking advantage of the strengths of each model, reducing training costs but still ensuring effective combination thanks to the shared Transformer-based architecture. However, there is a mismatch in the tokenization level between BARTpho-word (word level) and pre-trained GPT2 (byte level), which may negatively impact the decoder's performance (refer to Figure 5).

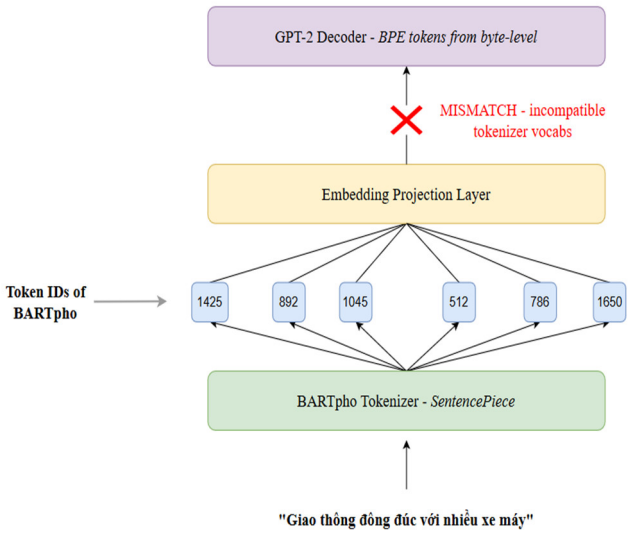

Figure 5. Illustrate of tokenizer-decoder mismatch when combining BARTpho with GPT2

Despite the mismatch between Tokenizer and Decoder, our model demonstrates strong performance in generating accurate and fluent Vietnamese captions. This observation can be attributed to two main factors.

First, although the token embeddings may not be fully compatible due to differences in tokenization scheme difference, much of the pre-trained structure of GPT-2 especially the multi-head self-attention layers remains intact and continues to play an important role in the model's language generation ability.To improve training efficiency and reduce the number of parameters to be tuned, Low-Rank Adaptation (LoRA) [16] is applied to the c_attn and c_proj layers in the self-attention and cross-attention modules of the decoder. Instead of adapting the full weight matrix $W \in R^{d \times k}$, LoRA freezes W and applies a trainable low-rank update:

$$W' = W + \Delta W = W + BA \quad (6)$$

Where $A \in R^{r \times k}, B \in R^{d \times r}$, with $r \ll \min(d, k)$. BA is a learnable low-rank matrix that approximates the weight update. W is the pre-trained (frozen) weight matrix (refer

to Figure 6). LoRA works by inserting low-rank matrices, which reduces the number of parameters that need to be updated, while keeping the pretrained backbone intact. This allows the model to retain most of the general knowledge learned from pretraining, reducing the impact of the new token split.

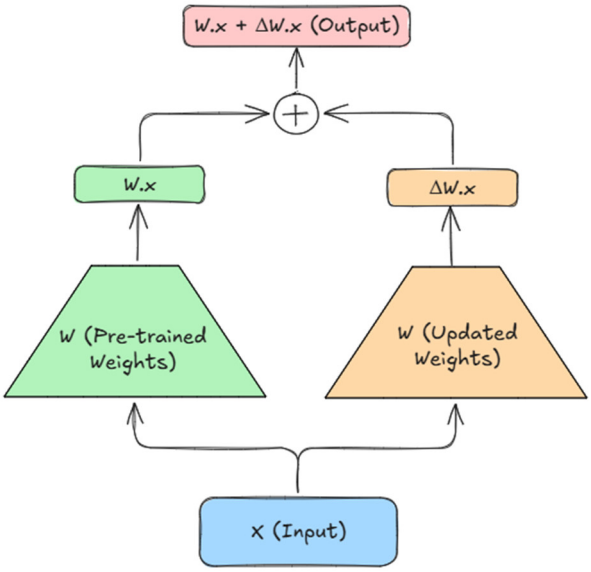

Figure 6. Illustrate of fine-tuning process with LoRA

Second, the Vietnamese caption dataset used for fine-tuning is of high quality. The captions have consistent syntax, using common traffic context words. This uniformity makes it easier for the decoder to learn the relationship between visual and linguistic information. Therefore, even if there is a mismatch in the token embedding due to differences between tokenizers, the decoder can still take advantage of the regularities in the caption structure to generate accurate and coherent Vietnamese sentences.

Table 3. PVG performance comparison on tpc37k with and without lora fine-tuning

Epochs	LoRA	BLEU-1	BLEU-4	ROUGE-L	CIDEr
2	✖	28.1	8.3	49.1	10.6
	✓	31.2	8.2	47.4	11.3
3	✖	28.2	8.0	49.1	10.4
	✓	32.4	10.2	48.8	19.2

Training on a dataset of more than 9,000 observations, when evaluating the two models, the following conclusions were drawn: training on all parameters, the numbers change very little over the epochs. When applying the LoRA method, the indices tend to increase steadily over the epochs, showing that LoRA is superior in updating important and necessary numbers. For that reason, the LoRA method is applied to the Image Captioning problem in this discussion paper.

3. RESULTS AND DISCUSSION

3.1. Quantitative Evaluation

We evaluate model performance on BLEU, ROUGE-L, and CIDEr. BLEU estimates n-gram precision but does not suit Vietnamese due to its flexible syntax and rich expression and has a tendency to penalize semantically correct captions. ROUGE-L values content recall and is suitable for evaluating whether essential information is retained, especially beneficial in structured, factual captions. CIDEr, which considers both term frequency and semantic saliency across multiple references, best reflects the quality of informative and context-sensitive descriptions. Since our model is interested in generating short, well-organized captions for assisting visually impaired users in perceiving traffic scenes, CIDEr and ROUGE-L are the most indicative evaluation metrics for the same.

Table 4. Data statistics amongst three datasets using for other models

Dataset	Images	Captions	Captions/ images	Avg. length
TPC37k [14]	37,056	9108	1:4	31.21
UIT-ViIC [12]	4,000	20,000	5:1	12.19
KTVIC [13]	4327	21,635	5:1	10.97

Table 5. Performance comparison of model evaluate measures using tpc37k

Model	Dataset	BLEU-1	BLEU-4	ROUGE-L	CIDEr
PVG	TPC37k	54.7	34.1	**84.7**	**142.5**
	KTVIC	36.9	12.7	41.9	98.4
	UIT-ViIC	42.3	16.5	47.5	96.5
ResNet101 LSTM [13]	TPC37K	31.3	13.1	44.6	33.3
	KTVIC	53.0	15.5	42.3	21.8
ViT - Transformer [13]	TPC37K	40.8	17.3	50.2	56.6
	KTVIC	**74.7**	**40.6**	59.7	136.0

In Table 5, the PVG model on TPC37K achieves ROUGE L(84.7) and CIDEr(142.5), outperforming the other configurations. These two metrics reflect the semantic matching and content richness; therefore, the high results indicate that PVG reproduces traffic information most completely and accurately. In contrast, ViT Transformer on KTVIC has the highest BLEU 1(74.7) and BLEU 4(40.6), indicating a very good surface-level n gram

matching ability but is still inferior to PVG in semantic depth (ROUGE L, CIDEr). This suggests that, in the highly structured traffic image description problem, PVG makes strong use of the specific grammatical information of TPC37K, while ViT Transformer focuses more on lexical coverage than content accuracy.

Table 6. Performance comparison of datasets using pvg model

Dataset	BLEU-1	BLEU-4	ROUGE-L	CIDEr
TPC37k	**54.7**	**34.1**	**84.7**	**142.5**
UIT-ViIC	42.3	16.5	47.5	96.5
KTVIC	36.9	12.7	41.9	98.4

Based on the comparative results displayed in Table VI, when trained on TPC37K, PVG achieved BLEU-434.1, ROUGE-L84.7 and CIDEr142.5, far surpassing the same model on UIT-ViIC and KTVIC. The difference shows that TPC37K has a highly compatible annotation structure (template) and traffic vocabulary with PVG's Transformer architecture, which helps the model learn fixed sentence templates and information fields (signs, lanes, safe travel directions, etc.) effectively. In contrast, UIT-ViIC and KTVIC contain free annotations with little structure, which reduces PVG's advantage. In summary, PVG demonstrates the highest compatibility with TPC37K's richly structured data, confirming the importance of grammatical structure and corpus features.

Briefly, the results prove the power of applying domain expertise and data augmentation methods in captioning images. PVG's performance on the TPC benchmarks demonstrates the power of adjusting model architecture and training data to the task-specific context requirements.

3.2. Generated Caption Analysis

As shown in Figure 7, the captions are extremely fluent and semantically sound, rightly captioning principal visual elements in Vietnamese traffic scenes. The model, on average, requires 6 to 8 seconds to output a caption for each image. Qualitatively, the predictions are concise, contextually accurate, and grammatically correct, with strong attention to principal traffic elements such as cars, road signs, pedestrians, and lane markings. In dense situations, the model shows good prioritization, frequently detecting important objects (e.g., pedestrians, motorbikes, buses) while retaining clear sentence coherence. This shows in dense scenes or in scenes with rich traffic activity. Although these are positive, there are some weaknesses. In specific cases, the model occasionally fails to detect fine-grained visual cues, such as traffic sign glyphs or subtle visual warnings. Additionally, directional information, particularly the "left"-"right" dichotomy is also subject to unreliability in certain situations, giving rise to the threat of ambiguity in spatial descriptions. These issues pose a need for enhanced spatial reasoning and observation of symbolic features in future revisions of models.

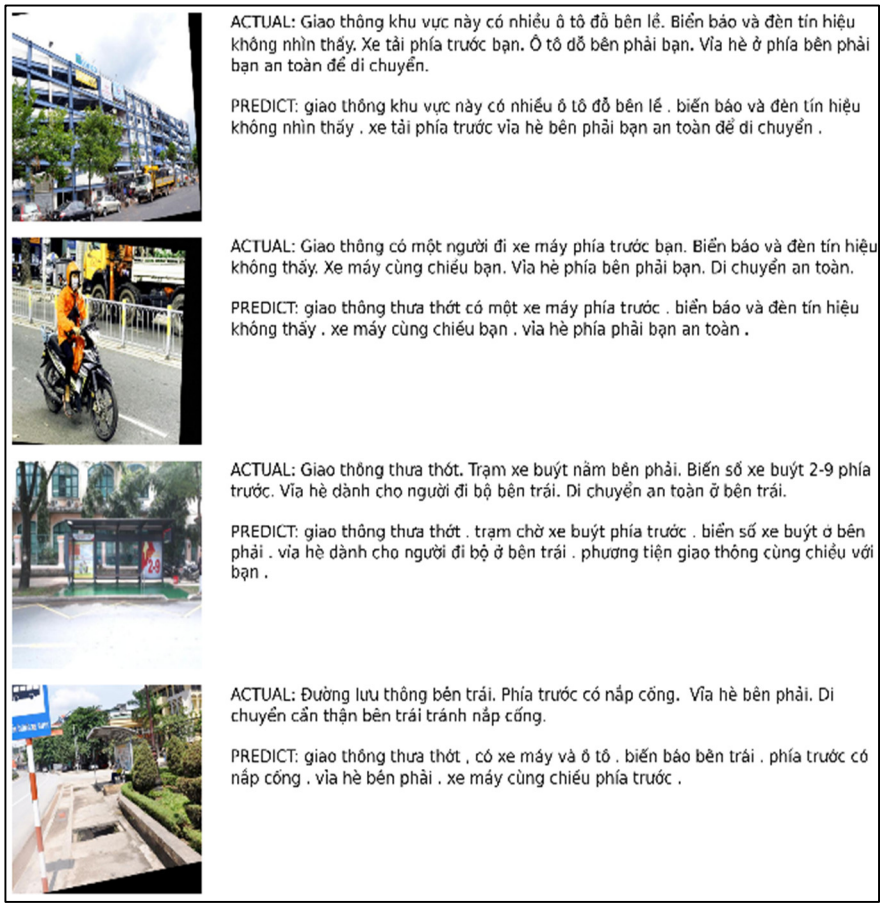

Figure 7. Examples of inference images from the TPC37k dataset with our proposed method (PVG)

4. CONCLUSION

This work introduces PVG, a multimodal image captioning model that combines ViT for visual encoding, GPT-2 for language decoding, and BARTpho for enhanced Vietnamese language comprehension. To address the tokenizer-decoder mismatch between BARTpho (SentencePiece) and GPT-2 (BPE), we utilize LoRA to

selectively fine-tune solely the final decoder layers, preserving pretrained knowledge with improved flexibility. Furthermore, captions on our specially formatted TPC37k dataset also follow a semantic template format that helps stabilize decoding and maintain relevance in the traffic scenario. Experiment results show PVG–Augment performs exceedingly well with 142.5 CIDEr and 84.7 ROUGE-L metrics, outperforming existing baselines. Results demonstrate the power of combining architecture-level adaptation, tokenizer alignment practices, and task-specific data structure.

REFERENCES

[1]. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Computer Vision and Pattern Recognition*, 2016.

[2]. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *Computer Vision and Pattern Recognition,* 2018.

[3]. L. Huang, W. Wang, J. Chen, X.Y. Wei, "Attention on Attention for Image Captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4634-4643, 2019.

[4]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017.

[5]. M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, WA, USA, p. 10578-10587, 2020.

[6]. Y. Pan, T. Yao, Y. Li, T. Mei, "X-Linear Attention Networks for Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[7]. S. Herdade, A. Kappeler, K. Boakye, J. Soares, "Image Captioning: Transforming Objects into Words," *Advances in Neural Information Processing Systems*, 32, 2019.

[8]. L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, J. Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA," in *Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition*, 13041-13049, 2019.

[9]. P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, "VinVL: Revisiting Visual Representations in Vision-Language Models," in *Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition,* 579-588, 2021.

[10]. B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," in *Conference on Computer Vision and Pattern Recognition*, 2023.

[11]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations,* 2021.

[12]. Q. H. Lam, Q. D. Le, V. K. Nguyen, N. Luu, T. Nguyen, "UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning," in *Computational Collective Intelligence*, 2020.

[13]. C. A. Pham, Q. V. Nguyen, H. T. Vuong, T. Q. Ha, "KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain," arXiv, 2024.

[14]. T. P. Dai, "Kaggle," 4 2025. [Online]. Available: https://www.kaggle.com/datasets/trieuphi/traffic-pictures-captioning.

[15]. N. L. Tran, D. M. Le, D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese," in *Interspeech*, 2022.

[16]. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations* (ICLR), 2023.

[17]. K. Nguyen, "Empirical study of feature extraction approachesfor image captioning in vietnamese," *Journal of Computer Science and Cybernetics*, 38, 327-346, 2022.

[18]. R. khan, M. S. Islam, K. Kanwal, M. Iqbal, M. I. Hossain, Z. Ye, "A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism," *Information Technology and Control,* 16, 2022.

[19]. A. M. Rinaldi, C. Russo, C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks," *Results in Engineering*, p. 14, 2023.

[20]. G. O. d. Santos, E. L. Colombini, S. Avila, "CIDEr-R: Robust Consensus-based Image Description Evaluation," in *Computer Vision and Pattern Recognition*, 2021.