JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Enhancing Semantic Coherence in Image Captioning via a Parameter-Efficient Refinement Framework

Huynh Anh Ngan Ha, Lam Hoang Phu Bui, Minh Sam Lam, Thi Dinh Nguyen*
Ho Chi Minh City University of Industry and Trade, Vietnam

*Corresponding author. *Email: dinhnt@huit.edu.vn*

ARTICLE INFO

Received: 06/04/2026
Revised: 12/05/2026
Accepted: 02/06/2026
Online First: 23/06/2026
Published:

KEYWORDS

Image Captioning;
Semantic Coherence;
Caption Refinement;
Parameter-Efficient Learning;
Two-Stage Framework.

ABSTRACT

Large-scale pre-trained visual-linguistic models have achieved significant progress in image annotation generation. However, the generated descriptions often suffer from limitations in semantic consistency, a lack of key image elements, and structural coherence. To overcome these limitations without requiring high-cost end-to-end refinement, this study proposes a two-stage parameter-efficient refinement framework. In the first stage, the pre-trained visual-linguistic model is fixed to generate initial annotations from the input image. In the second stage, the problem is redefined as a conditional text generation task, where the pre-trained linguistic model is adjusted using low-order adaptive techniques to improve grammatical structure and enhance semantic coherence, while preserving previously learned knowledge. Experimental results on the Flickr30k dataset, using BLEU-n and METEOR scales, demonstrate that the proposed method significantly enhances the quality of expression and semantic consistency compared to the baseline model, while maintaining a low number of trained parameters. The proposed framework offers a cost-effective solution for enhancing the quality of semantically oriented annotations, and also lays the groundwork for further research on fine-tuning efficient parameters in multimodal language generation models.

Doi: https://doi.org/10.54644/jte.2026.2126

1. Introduction

Image captioning is a fundamental problem combining computer vision and natural language processing, aiming to generate descriptions of image content in natural language. In recent years, the development of large-scale pre-trained linguistic vision models has brought about improvements in accuracy on scales such as BLEU-n, METEOR, or CIDEr [1]. However, despite achieving high scores, many systems still have limitations in semantic coherence. The generated captions may lack coherence between sentence components, inaccurately describe the relationships between objects, or exhibit hallucination. Most current approaches focus on extending the model architecture or fine-tuning all parameters end-to-end, which leads to high computational costs and limits their implementation in resource-limited environments [2].

To address semantic consistency limitations, this paper proposes a two-stage, parameter-optimized refinement framework that does not require updating the entire linguistic vision model. In the first stage, a pre-trained linguistic vision model generates initial annotations from the input image, with its parameters remaining fixed to preserve learned multimodal knowledge. In the second stage, the initial annotation is then redefined as input to the conditional text generation problem. Here, a pre-trained linguistic model is refined using a parameter-efficient mechanism, with only some parameters updated instead of retraining the full network. This refinement stage optimizes syntactic structure, adjusts entity relationships, and enhances semantic coherence between sentence components. This two-tier design separates visual information extraction from linguistic expression correction, improving annotation quality without significantly increasing training costs or requiring large computational resources.

The main contributions of the paper include: (1) Proposing a two-stage refinement framework to improve semantic consistency without retraining the visual-linguistic model; (2) Redefining the

annotation improvement problem as a conditional text generation task, thereby separating the visual extraction and linguistic refinement processes; (3) Applying an efficient parameter refinement strategy to reduce training costs and suit a resource-constrained environment; (4) Experiments on the Flickr30k image set [3] and error analysis show that the method improves both quantitative scale and coherence in expression.

2. Related works

Many different approaches have been proposed to improve the content accuracy and semantic coherence of captions, from traditional CNN and RNN models, attention mechanisms, Transformers, to external knowledge integration methods such as scene graphs and Knowledge Graphs [4]–[7]. However, each approach has certain limitations related to the semantic representation of images, computational cost, or the complexity of the system. This section will synthesize and analyze typical groups of works.

Early research in image captioning primarily focused on improving encoder–decoder architectures, including CNN–RNN frameworks and Transformer-based models, to enhance visual representation and sequence generation performance. The study entitled "Performance Evaluation of CNN-Based Encoders for Image Captioning" investigated the effectiveness of CNN-based encoder architectures in the image captioning task. The authors compared several widely used backbone networks, including ResNet, ConvNeXt, and Swin Transformer, within a traditional encoder–decoder framework to assess their impact on feature representation and caption generation quality [4]. The results showed that Transformer-based architectures like Swin Transformer have better global context capture than pure CNN, thereby improving the quality of feature representation. The advantage of this method is the analysis of the role of the encoder in the caption generation process. However, the work primarily focuses on improving visual representation without deeply addressing the issues of linguistic coherence or semantic consistency at the text generation level; therefore, instances of missing objects and misinterpretations still exist in complex scenes. The study entitled "Exploring Visual Relationships via Transformer-Based Graphs for Enhanced Image Captioning" proposed integrating relational information between image regions to enhance contextual understanding during caption generation [5]. The advantage of this method is that it emphasizes the role of visual relationships in improving semantic accuracy. However, modeling relationships depends on the quality of image region detection and increases computational complexity.

A knowledge-enhanced framework was proposed in which entities extracted from images are mapped to an external knowledge graph, and the corresponding knowledge embeddings are incorporated into the Transformer attention mechanism to improve caption generation performance [6]. In the paper, the method extracts entities from images, maps them to an external knowledge graph, and then inserts knowledge embeddings into the attention layer of the Transformer to support annotation generation. However, the method requires a complex entity linking process and depends on the completeness of the Knowledge Graph. A knowledge graph–augmented framework was introduced to enhance image captioning by enabling the model to access external knowledge sources rather than storing all information within model parameters, thereby extending its semantic reasoning capability [7]. This study highlights the effectiveness of KG augmentation when combined with vision-language pretraining to reduce data and parameter requirements. The limitations are that the method is heavily dependent on the quality of the mapping between KG and the captioning model.

An ensemble model combining CNN and Transformer encoder–decoder architectures with an integrated attention mechanism was proposed to improve caption generation performance [8]. In this approach, input images are first encoded by a CNN to extract spatial features, which are then passed to a Transformer decoder that dynamically focuses on salient image regions at each generation step [8]. However, the method requires high computational costs due to the need to train and infer many parallel networks, limiting its practical implementation and failing to fully address the semantic coherence issue. A number-controlled captioning network was introduced to regulate both the quantity and diversity of captions generated for each image, thereby enhancing output controllability and variation [9]. Instead of generating only a single caption, NumCap assigns a control token to the training input, allowing the model to learn how to generate different captions corresponding to each numerical value. The main advantage of this research is the expansion of output control capabilities and the limitation of sentence

pattern repetition. However, the approach primarily focuses on content diversification rather than improving coherence or addressing issues of missing objects and misinterpretation, thus remaining limited in terms of semantic consistency in complex scenarios.

To enhance semantic understanding beyond visual features, recent studies have incorporated external knowledge sources such as scene graphs and knowledge graphs into the caption generation process. A novel architecture was proposed that integrates external commonsense knowledge from ConceptNet into the caption generation process to mitigate mismatches between textual descriptions and visual ambiguity in images [10]. The method uses CLIP-based image encoders (ViT) and Faster R-CNN to extract global information and object regions, then performs multi-step cross-attention alignment between image and text features to optimize cross-modal consistency. The highlight is the entity linking process that maps object features to the knowledge graph, then uses TransE embedding and multi-level reasoning to add external knowledge to the GPT-2 decoder. Experiments on MS-COCO and Flickr30k showed a significant improvement in the CIDEr score (142.6 on MS-COCO and 78.4 on Flickr30k). The advantages of this study include the successful integration of knowledge from ConceptNet to increase semantic richness and accuracy in annotations, and the limitation of methods that rely on the quality of entity linking results and require extensive computation for retrieving external knowledge. Graph-based Image Captioning with Semantic and Spatial Features (2025) [11] focuses on exploiting semantic and spatial graphs between objects to improve caption context. The resulting graphs are integrated with CNNs and word embeddings into the LSTM decoder via multimodal attention, allowing the model to generate captions with more contextual information. The advantage is that the model combines spatial and semantic features to reduce the phenomenon of missing objects or context; however, the method still uses the old LSTM network and depends on the quality of the GCN built from RelTR, which may limit its scalability to modern Transformer architectures. A method was proposed that replaces traditional CNN encoders with Vision Transformers while integrating commonsense knowledge graphs to provide supplementary contextual information for caption generation, thereby enhancing overall scene understanding [12]. Experiments on the published standard benchmark show small improvements for the BLEU-1, BLEU-4, METEOR, ROUGE-L, and CIDEr measures. The advantage is the approach of combining modern Transformer architecture with KG to enhance the ability to model semantic dependencies; the limitation is that the improvement is not large, and the model does not deeply exploit the complexities of KG, such as multi-hop reasoning or rich triplets.

Given this complexity, this study chooses a parameter-based optimization (LoRA) approach on the pre-training model, instead of designing complex KGs. This approach helps keep the model lightweight, reduces training costs, and enhances semantic representation by leveraging existing cognitive knowledge within the model, aiming to improve the coherence and consistency of annotations in a practical and scalable way.

3. Methods and proposed model

One of the core challenges of image captioning today is no longer the ability to recognize individual objects, but rather ensuring semantic coherence and the ability to organize information contextually within the entire description. Recent approaches often address this by integrating Knowledge Graphs or scene graphs into the processing workflow to supplement relationships and external knowledge. However, this approach significantly increases system complexity, requiring many intermediate modules (object detection, entity linking, graph encoding, fusion mechanism), and also increases training and inference costs. Based on this, the proposed model selects a strategy for efficient parameter tuning to achieve a balance between enhancing semantic representation and controlling model complexity.

3.1. Proposed overall system architecture

The proposed architecture (Figure 1) inherits the existing vision language backbone, including a Vision Encoder and a Language Decoder. This proposed architecture consists of the following stages:

Stage 1: Use the BLIP model [13] to create an initial description; then compare it with the ground truth to synthesize the input caption for training the T5 model [14].

Stage 2: With a new input image, BLIP creates a rough description, then transmits the rough description to the trained T5 model to adjust the grammatical structure.

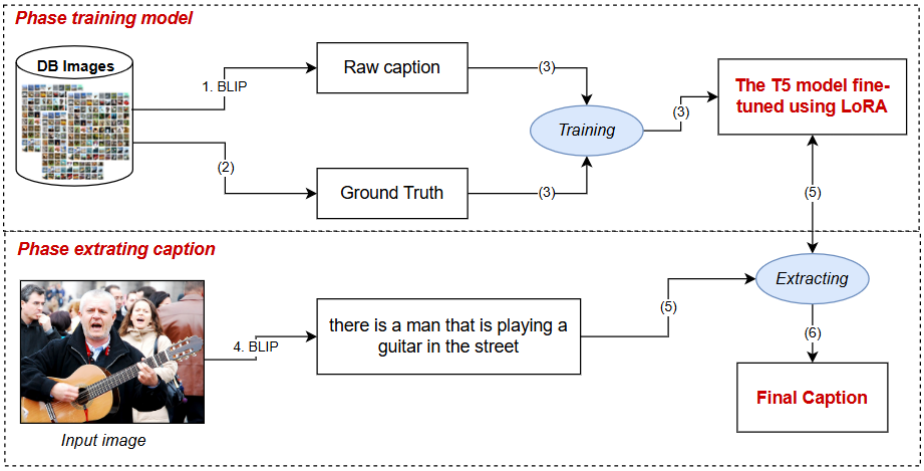


Figure 1. *Proposed model for extracting image caption.*

3.2. Parameter Constrained Optimization Problem

Pre-trained vision language models often have a very large number of parameters. If the entire model were fine-tuned for the image captioning problem, the computational cost and memory requirements would increase significantly, and the risk of overfitting would also increase if the training data were insufficient. Therefore, instead of updating all parameters of the original model, this study treats the fine-tuning process as an optimization problem with a constraint on the number of parameters. Specifically, the initial pre-training parameters are kept fixed, and only a small portion of the additional parameters are allowed to be learned. The goal is to improve the quality of the generated annotations while ensuring a compact and stable model.

To accomplish this, the study applies the Low-Rank Adaptation (LoRA) method [15]. LoRA adds a low-rank adaptation component to the original weights.

$$W' = W + BA \tag{1}$$

In this study, the LoRA configuration hyperparameters were set with a rank value of r = 8 and a scaling factor of alpha = 16. To enhance reproducibility and optimize parameter usage efficiency for the Transformer architecture of the T5 model in stage 2, low-rank adaptation matrices were directly integrated into the attention layers. This configuration was applied to the Query projection (Wq) and Value projection (Wv) matrices within both the self-attention and cross-attention blocks of the Encoder and Decoder layers. Selective manipulation of the attention layers allows the T5 model to focus on optimizing its ability to capture dynamic semantic relationships between linguistic entities from the raw text sequence of stage 1.

In this approach, A and B are low-rank matrices with much smaller dimensions than W, and only the parameters of these two matrices are optimized during training. As a result, the number of parameters to be updated is significantly reduced compared to full fine-tuning [15].

This approach helps the model learn the necessary adjustments to enhance semantic consistency and linguistic expressiveness, while preserving the knowledge learned from the pre-training phase. Therefore, LoRA acts as an efficient fine-tuning mechanism, suitable for the goal of improving caption quality without significantly increasing system complexity.

3.3. LoRA-Based Training and Optimization

Based on the parameter-constrained optimization problem presented in Section 3.2, the model training process updates only the adjustment parameters θ (LoRA), while keeping the original

parameters θ of the pre-training model fixed. This approach reduces the number of parameters to be optimized while still allowing the model to adapt to the data of the image captioning problem [16].

For each input image I, the visual encoder extracts high-level features and transmits them to the language generator. The captioning process takes place according to an autoregressive mechanism: at each time step t, the probability of the next word is calculated based on the image features and the sequence of previously generated words. The model is trained by maximizing the probability of the correct caption sequence in the training dataset [16]. In terms of optimization, the LoRA parameters are updated using a gradient-based algorithm. Because only a small fraction of the parameters are trained, the number of variables requiring tuning is significantly reduced compared to full fine-tuning, resulting in faster, more stable convergence that is less sensitive to data noise. This aligns with building a computationally efficient model while maintaining the semantic quality of the generated annotations.

4. Evaluation of experimental results

4.1. Data and experimental environment

The experimental configuration used a computer with an Intel Core i7-2600K CPU (4 cores, 8 threads, 3.4–3.8 GHz) and a GTX 1050 Ti GPU (4 GB VRAM). This configuration was insufficient for training large models, but performed well in inference from BLIP and local data management. The system was developed in Python 3.10+, using PyTorch for deep learning, Transformers for BLIP and T5, and PEFT for the LoRA mechanism. T5 training and LoRA refinement were performed on Google Colaboratory with an NVIDIA Tesla T4 GPU (16 GB VRAM), which supports mixed precision computing and is capable of handling Transformer models with batch sizes of 16–32.

The Flickr30k experimental data, with 31,784 images, has sufficient properties to study and emphasize the relationship between entities in images, which is very suitable for studying image annotation problems. The data is divided into three sets: train (29,783 images), validation set (1,000 images), and test set (1,000 images) [3]. To evaluate efficiency in resource-constrained environments, the proposed LoRA-based method was compared against Full Fine-Tuning using the T5 architecture. Key hardware utilization metrics are summarized in Table 1.

Table 1. *Computational resource usage between Full Fine-Tuning and the Proposed Method.*

Tuning Strategy	Trainable Parameters	Training VRAM (GB)	Training Time (ms/Epoch)
Full Fine-Tuning	220 M	14.2	45.05
Proposed method (LoRA)	1.8 M	5.4	15.09

4.2. Experimental results

This illustrates the experimental results of extracting captions from an input image.

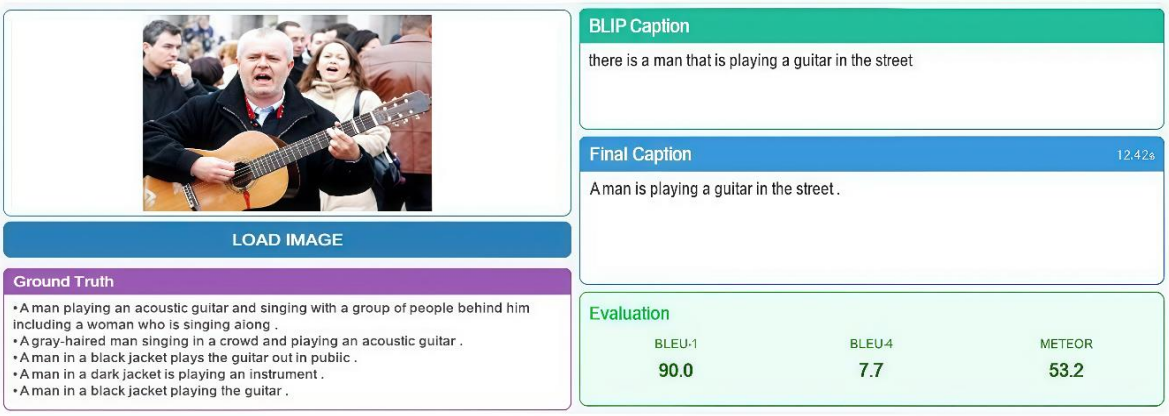
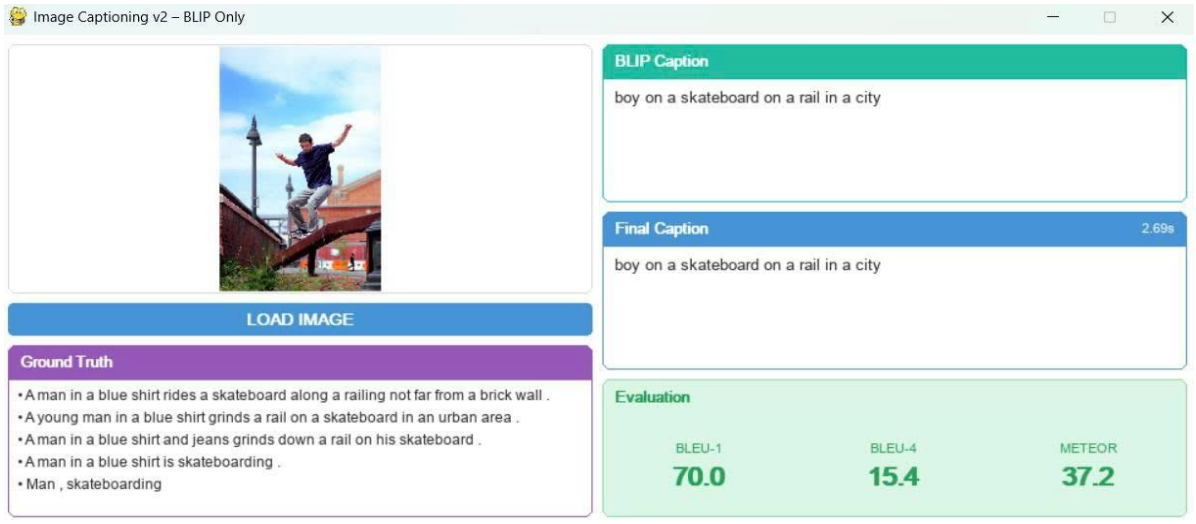
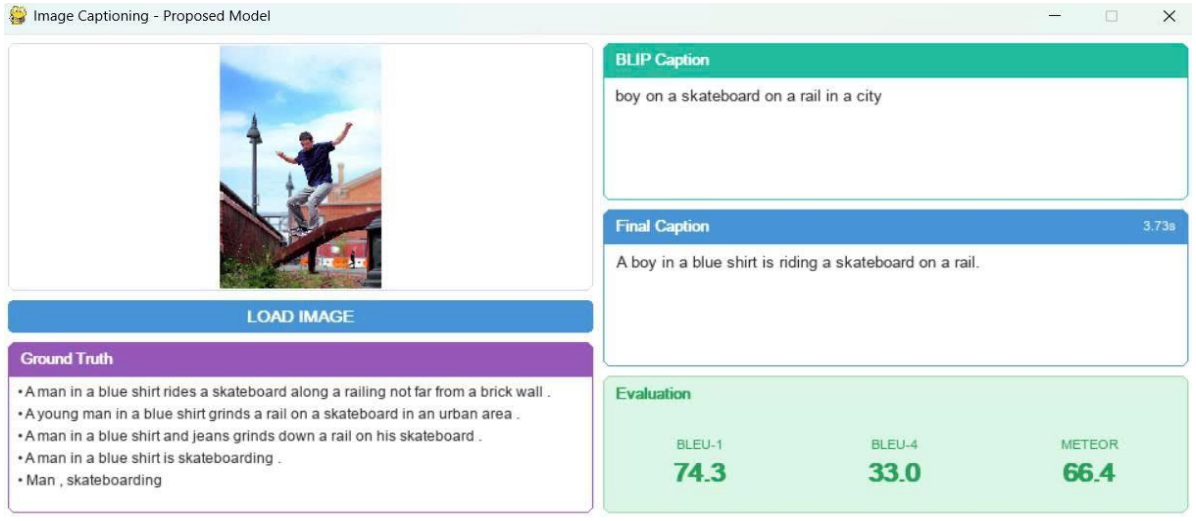


Figure 2. *Image caption results.*

The experiment illustrates an integrated model to demonstrate the entire image annotation process and evaluate the output (Figure 2). The system allows the user to load an initial image, then sequentially displays the components related to the processing and annotation results. The application displays the entire ground truth as a reference for evaluating the quality of the generated annotation. Next, the system displays the raw annotation generated by the BLIP model and the final annotation after optimization by the T5 model. Based on the comparison between the generated annotation and the ground truth annotations, automatic evaluation scales such as BLEU-1, BLEU-4, and METEOR are calculated and displayed at the bottom of the screen.



(a)



(b)

Figure 3. *Image caption results: (a) BLIP caption only; (b) Proposal Model.*

The comparative experimental results (Figure 3) show that the BLIP-only model in Figure 3a produces the caption “boy on a skateboard on a rail in a city”. While Figure 3b produces the caption “A boy in a blue shirt is riding a skateboard on a rail”. This result demonstrates the effectiveness of the proposed experimental model.

Based on the experimental results obtained, the proposed model shows the potential to improve the quality of image annotations compared to the initial raw descriptions generated by the visual model. This improvement is demonstrated not only in the degree of similarity to ground truth annotations but also in the stability of sentence structure and semantic expressiveness.

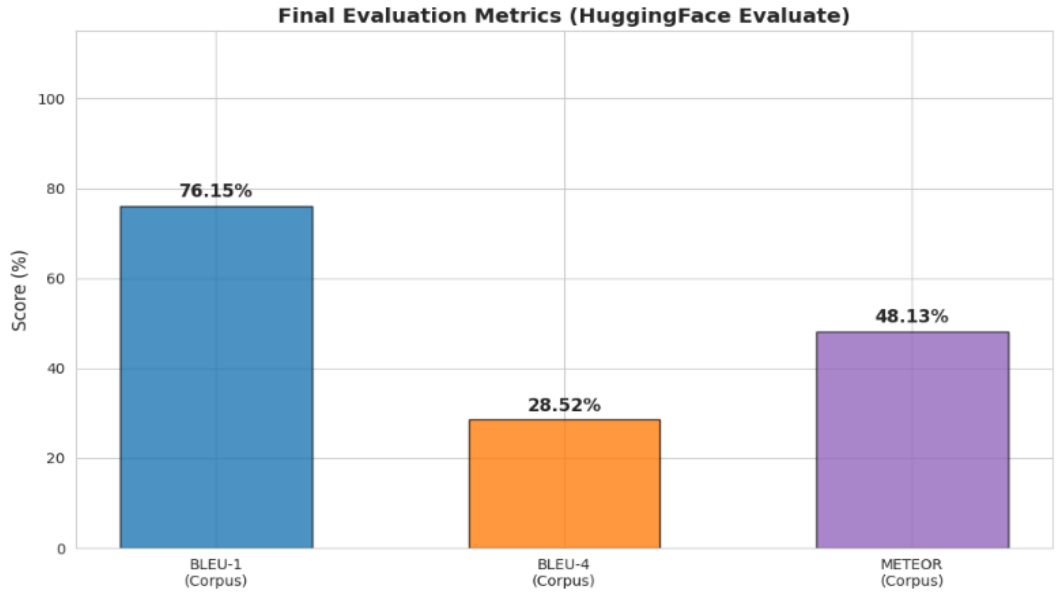
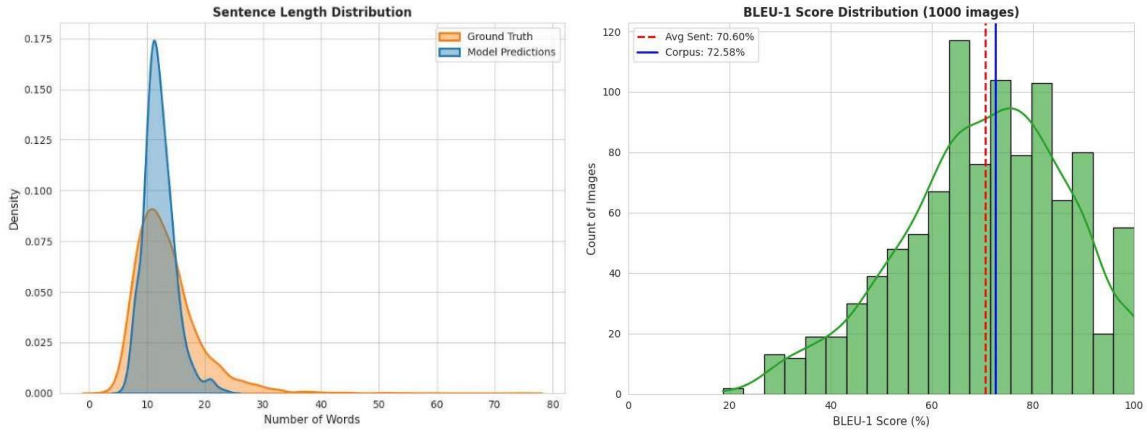


Figure 4. *Summary chart of evaluation indicators.*

The summary chart of evaluation indices shows that the BLEU-1 model achieved 76.24, the BLEU-4 model achieved 28.70, and the METEOR model achieved 48.08 (Figure 4). The high BLEU-1 score indicates the model's ability to accurately represent keywords and main content in ground truth annotations, while BLEU-4 and METEOR demonstrate that the model has acquired sentence structure and semantic fluency at a fairly high level. This shows the clear role of the T5 model in improving the coherence, grammar, and naturalness of the final descriptive sentence.



(a) *(b)*

Figure 5. *Evaluation charts: (a) Sentence length distribution chart; (b) BLEU-1 score distribution chart for each image.*

The sentence length distribution chart shows that the sentences generated by the model are mostly between 10 and 15 words, relatively consistent with the distribution of ground truth annotations (Figure 5a). However, the model's distribution tends to be narrower and less frequent with longer sentences, indicating that the model prioritizes concise descriptions while still covering the main content of the image. This reflects the T5 model's ability to control sentence length and structure well after being fine-tuned with LoRA.

Considering the BLEU-1 score distribution by image, most samples scored in the medium to high range, with an average sentence score of approximately 70.60 and a corpus-level BLEU-1 score of 76.24 (Figure 5b). The small difference between these two values indicates that the model performs stably across the entire test set, without being overly dependent on a few individual samples. The score

distribution is concentrated, with few outliers in the low-score region, suggesting that the model has relatively good generalizability across various image contexts.

4.3. Analyzing, comparing and evaluating experimental results

To further clarify the extent of the proposed model's contribution, we compared and evaluated it with the original baseline model; the results are presented in Table 2.

Table 2. *Performance evaluation results of the proposed model on Flickr30k.*

Model	BLEU-1	BLEU-4	METEOR
BLIP [17]	60.69	18.20	44.22
Proposed model	72.58	28.70	48.08

From the results obtained from the two models in Table 1, we see that the proposed model model successfully achieved consistent improvement across all rating scales compared to the original BLIP model. Specifically, the BLEU-1 score increased from 60.69 to 76.24, indicating a significant improvement in lexical match between the generated annotation and the ground truth. Simultaneously, the BLEU-4 score increased from 18.20 to 28.70, reflecting the ability to generate longer word sequences with more logical grammatical and contextual matching, rather than just matching at the single word level. In addition, the METEOR index also increased from 44.22 to 48.08, showing that the model not only improved in form but also better represented semantic similarity and synonym selection compared to the human-compiled annotation.

Simultaneously, to place the model within the broader context of recent research directions, we extended the comparison to several representative methods published in recent years. The corresponding results are summarized in Table 3. However, to clarify the focus of the model according to the scales, it is necessary to demonstrate the empirical results obtained compared with previous methods and models on the same dataset.

Table 3. *Results of generating annotations on the Flickr30k set.*

Model	BLEU-1	BLEU-4	METEOR
Kalimuthu et al. (2021) [18]	64.7	22.4	19.7
LSTNet (2023) [19]	67.1	23.3	20.4
Trans-KG (2021) [20]	67.6	26.0	21.9
NumCap (2023) [9]	69.4	25.4	25.1
MSCI (2021) [21]	70.5	29.4	23.8
Proposed model	72.58	26.48	28.68

The results in Table 2 show that the proposed model achieves higher performance than many typical image annotation generation methods published in recent years on the same Flickr30k dataset. Specifically, BLEU-1 scores are higher than Kalimuthu et al. (2021), LSTNet (2023), Trans-KG (2021), and NumCap (2023) by approximately 2.1–7.9, indicating a significant improvement in vocabulary coverage and fit between generated sentences and standard annotations in BLEU-1s. On the BLEU-4 scale, the model demonstrates the ability to maintain long-term semantic structure at a level competitive with knowledge mining or semantic relation methods such as Trans-KG and NumCap, while being only lower than MSCI (2021) 2.9, which uses multi-layered semantic context during annotation generation. In particular, the METEOR score significantly outperformed all other models, reaching 48.7, while the remaining models ranged from 19 to 25.1. This clearly reflects the ability to generate descriptive sentences with high naturalness, coherent linguistic structure, and a semantic similarity close to human expression. These results demonstrate that the approach of refining and standardizing annotations through the T5 language model, based on raw output from BLIP, yields noticeable improvements in language quality without significantly increasing inference costs.

Knowledge Graph integration methods often require complex construction, including entity linking, graph construction, and reasoning modules. These components significantly increase the system's complexity and inference cost. In contrast, the method proposed in this study does not alter the model's core architecture and does not add external knowledge. Instead, knowledge is adjusted through a purposeful parameter fine-tuning mechanism, keeping the system lightweight while improving semantic representation. Despite achieving positive results, the current study has some limitations. Firstly, the experiment was only evaluated on a few standard datasets and has not been extended to other data domains. Secondly, the study has not analyzed in detail the inference time and memory consumption in a real-world implementation environment.

Experimental results and error analysis on 1,000 samples using the Sentence-BERT method [22] are shown in Figure 6.

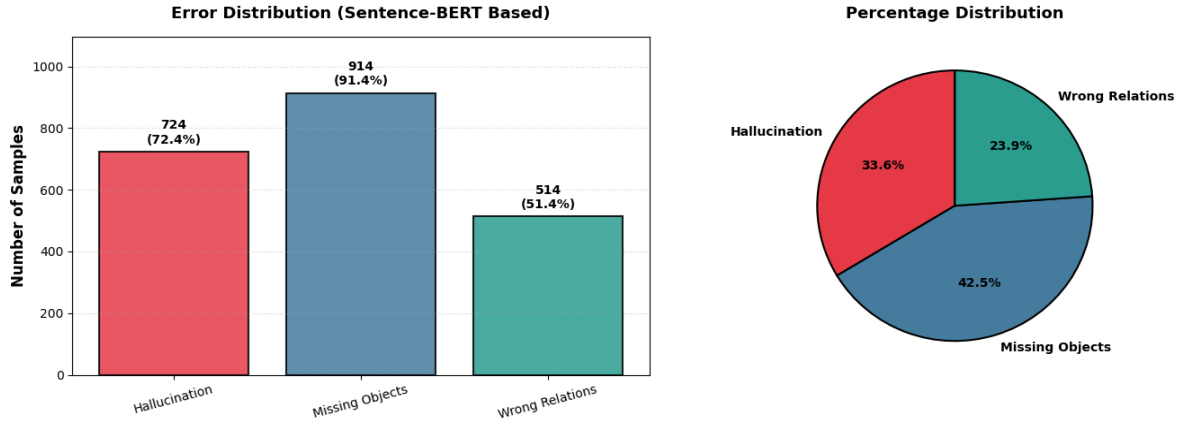


Figure 6. *Error distribution chart and error proportion chart.*

Based on the error distribution graph (Figure 6), the Missing Objects error accounts for the highest percentage with 914 samples (42.5%), indicating that the model often misses important entities in the image when generating annotations. This is followed by the Hallucination error with 724 samples (33.6%), reflecting the phenomenon of the model inferring additional objects or details that do not exist in the image. The Wrong Relations error accounts for the lowest percentage with 514 samples (23.9%); however, it still shows that the model sometimes does not accurately describe the relationships between entities. This result suggests that the main challenge of the system lies not in sentence structure, but in the level of full coverage of visual content and control of semantic inference, thus suggesting improvements focusing on enhancing entity recognition and contextual constraints.

The experimental results consistently demonstrate that the proposed parameter-efficient semantic enhancement strategy achieves competitive or superior performance compared to full fine-tuning, while dramatically reducing the number of trainable parameters. This confirms that semantic coherence can be improved through targeted adaptation mechanisms (LoRA and bias tuning), without relying on heavy knowledge graph construction or complex external reasoning pipelines.

5. Conclusion and future developments

This study proposes an improved framework to enhance semantic coherence in image captioning. The method utilizes a pre-trained visual-linguistic model to generate initial captions, and then applies a linguistic model with efficient parameter refinement (LoRA) to enhance the grammatical structure and semantic coherence of the sentences. Experimental results on the Flickr30k dataset show that the proposed model achieves significant improvements in metrics such as BLEU-1, BLEU-4, and METEOR compared to other models, while maintaining low training costs due to the small number of parameters that require updates. This demonstrates that separating the visual information extraction process from the parameter refinement process of the linguistic model is an effective approach.

However, the study still has some limitations, such as the model not fully utilizing information about entities in the image, which can lead to missing objects or incorrect inferences of relationships between

objects, and the experiments being conducted only on a standard dataset. In the future, the study could expand its evaluation to multiple datasets and integrate additional external knowledge, such as Knowledge Graphs or relationship recognition mechanisms, to improve semantic representation and reduce errors in the annotation generation process.

Furthermore, error analysis revealed that "Missing Object" remains a prominent issue (42.5%), as the text-only Stage 2 filter fails to recover image details missed in Stage 1. To mitigate this, our next work will focus on integrating multimodal clues, specifically by implementing Object Hints or directly inserting Image Features into the T5 system to provide a more robust image base.

Acknowledgments

The authors would like to thank the Faculty of Information Technology, Ho Chi Minh City University of Industry and Trade which is sponsor of this research. We also thank anonymous reviewers for their helpful comments on this paper.

Conflict of Interest

The authors declare no conflict of interest in this article.

REFERENCES

[1] J. Li, N. Xu, W. Nie, and S. Zhang, "Image captioning with multi-level similarity-guided semantic matching," *Vis. Informatics*, vol. 5, no. 4, pp. 41–48, 2021.
[2] J. C. Hu, R. Cavicchioli, and A. Capotondi, "Exploiting multiple sequence lengths in fast end-to-end training for image captioning," in *Proc. IEEE Int. Conf. Big Data (BigData)*, 2023, pp. 2173–2182.
[3] R. Muzaffar, S. Y. Arafat, J. Rashid, J. Kim, and U. Naseem, "UICD: A new dataset and approach for Urdu image captioning," *PLOS ONE*, vol. 20, no. 6, p. e0320701, 2025.
[4] A. C. Hoang, D. C. Nguyen, and H. L. Nguyen, "Performance evaluation of CNN-based encoders for image captioning," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, 2023, pp. 212–217.
[5] J. Li, Z. Mao, H. Li, W. Chen, and Y. Zhang, "Exploring visual relationships via transformer-based graphs for enhanced image captioning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 5, pp. 1–23, 2024.
[6] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," *Pattern Recognit. Lett.*, vol. 143, pp. 43–49, 2021.
[7] S. S. Santiesteban, S. Atito, M. Awais, Y. Z. Song, and J. Kittler, "Improved image captioning via knowledge graph-augmented models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 4290–4294.
[8] I. Al Badarneh, B. H. Hammo, and O. Al-Kadi, "An ensemble model with attention-based mechanism for image captioning," *Comput. Electr. Eng.*, vol. 123, p. 110077, 2025.
[9] A. Abdussalam, Z. Ye, A. Hawbani, M. Al-Qatf, and R. Khan, "NumCap: A number-controlled multi-caption image captioning network," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 4, pp. 1–24, 2023.
[10] L. Wang, M. Jiao, Z. Li, M. Zhang, H. Wei, and Y. Ma, "Image captioning model based on multi-step cross-attention cross-modal alignment and external commonsense knowledge augmentation," *Electronics*, vol. 14, no. 16, p. 3325, 2025.
[11] M. J. Parseh and S. Ghadiri, "Graph-based image captioning with semantic and spatial features," *Signal Process. Image Commun.*, vol. 133, p. 117273, 2025.
[12] Y. A. Thakare, K. H. Walse, M. Atique, and V. M. Thakare, "Insightful analysis of image captioning models with Image Captions100," *AIP Conf. Proc.*, vol. 3327, no. 1, p. 020010, 2025.
[13] M. Limbu and D. Banerjee, "MedBLIP: Fine-tuning BLIP for medical image captioning," *arXiv:2505.14726, 2025*.
[14] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer-based model for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, pp. 2585–2594, 2022.
[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, and S. Wang, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
[16] M. A. A. Khan, Z. U. Rehman, J. Ma, and H. Ma, "Optimization of LoRa for BIoT based on ML: A case of ESL," *Alex. Eng. J.*, vol. 85, pp. 185–206, 2023.
[17] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 12888–12900.
[18] M. Kalimuthu, A. Mogadala, M. Mosbach, and D. Klakow, "Fusion models for improved image captioning," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 381–395.
[19] Y. Ma, J. Ji, X. Sun, Y. Zhou, and R. Ji, "Towards local visual modeling for image captioning," *Pattern Recognit.*, vol. 138, p. 109420, 2023.
[20] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "Exploring region relationships implicitly: Image captioning with visual relationship attention," *Image Vis. Comput.*, vol. 109, p. 104146, 2021.
[21] H. Tian, H. Mo, and L. Jiang, "Image caption generation using multi-level semantic context information," *Symmetry*, vol. 13, no. 7, p. 1184, 2021.
[22] J. Seo, S. Lee, L. Liu, and W. Choi, "TA-SBERT: Token attention sentence-BERT for improving sentence representation," *IEEE Access*, vol. 10, pp. 39119–39128, 2022.

Huynh Anh Ngan Ha is currently pursuing a Bachelor's degree in Software Engineering at Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam, expected to graduate in 2026. This month, she started an internship as a Business Analyst at UNIT while still being a student at the Faculty of Information Technology, Ho Chi Minh City University of Industry and Trade, Vietnam. Her research focuses on Software Engineering, Deep Learning and Business Process.

Email: hahuynhanhngan@gmail.com. ORCID: https://orcid.org/0009-0003-5613-4630

Lam Hoang Phu Bui is currently pursuing a Bachelor's degree in Software Engineering at Ho Chi Minh City University of Industry and Trade, Ho Chi Minh City, Vietnam, expected to graduate in 2026. He is currently working as a freelance AI Engineer. His research focuses on Software Engineering, Deep Learning and Automation.

Email: hoangphu130404@gmail.com. ORCID: https://orcid.org/0009-0005-2651-7053

Minh Sam Lam expected to receive my Bachelor's degree in Software Engineering from Ho Chi Minh City University of Industry and Trade (HUIT) in 2026. His core research interests include Software Engineering and Deep Learning, with practical experience in developing language models and data processing. He is currently focusing on optimizing software development lifecycles through advanced machine learning algorithms.

Email: lamminhsam123@gmail.com. ORCID: https://orcid.org/0009-0000-3951-4699

Thi Dinh Nguyen graduated in Pedagogy Informatics Ho Chi Minh City University of Education in 2006, and received a Master's degree in industry Data transmission and computer network at Ho Chi Minh City Institute of Post and Telecommunications Technology Ho Chi Minh City in 2011. In 2023, she received a PhD degree in Computer Science from the University of Science, Hue, Vietnam. Her field research includes image processing, image retrieval, and information system.

Email: dinhnt@huit.edu.vn. ORCID: https://orcid.org/0000-0003-3428-3101