PhantomTroupe at ImageEval 2025 Shared Task: Multimodal Arabic Image Captioning through Translation-Based Fine-Tuning of LLM Models

Muhammad Abu Horaira*, Farhan Amin*, Sakibul Hasan*,
Md. Tanvir Ahammed Shawon, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004029, u2004068, u2004043}@student.cuet.ac.bd

Abstract

Generating culturally accurate captions for images in Arabic remains a challenging task due to the language's rich morphology, complex syntax, and diverse cultural contexts.Cultural preservation involves capturing the significance and emotional resonance of images related to Palestinian heritage, ensuring accurate representation for future generations. We present a translation-assisted, instruction-tuned multimodal pipeline for Arabic image captioning, developed for the ImageEval 2025 Shared Task Subtask 2: on the evaluation of image captioning models.Our approach leverages the Qwen2.5-VL-7B-Instruct model with 4-bit quantization, fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) with LoRA. We implemented a pipeline involving translation of Arabic captions to English, followed by back-translation to generate fluent Arabic outputs. We evaluated several vision-language models, including Qwen2.5 VL (7B), Llama 3.2 (11B), and Pixtral (12B). The Qwen2.5 VL (7B) model achieved a BLEU-1 score of 22.6, a Cosine Similarity of 57.48, and an LLM Judge Score of 31.43, securing third place in the competition. These results underscore the potential of instruction-tuned multimodal models to produce culturally sensitive Arabic captions.

1 Introduction

Image captioning, the task of generating natural language descriptions for visual content, has advanced rapidly with the rise of deep learning and vision-language models. While these techniques have achieved impressive results in English and other resource-rich languages, extending them to Arabic presents distinctive challenges due to the language's morphological richness, syntactic flexibility, and dialect diversity(Al-Khalifa et al., 2021). Arabic words often encode extensive grammatical information within a single token, variations in word order can influence fluency, and regional dialects differ both lexically and culturally. The ImageEval 2025, First Arabic Image Captioning Shared Task addresses these gaps by introducing the first open-source, manually captioned Arabic dataset, enabling the development of culturally relevant models (Bashiti et al., 2025). In this paper, we present the Phantom Troupe team's participation in Subtask-2,where we developed a translation-assisted, instruction-tuned multimodal pipeline using the Qwen2.5-VL (7B) model.

Our key contributions include:

- We Developed a bidirectional translation pipeline that significantly improved Arabic image captioning by leveraging multilingual pretraining data, producing more accurate and contextually relevant captions.

- We Analyzed the effects of preprocessing techniques (e.g., RGB versus grayscale inputs), translation quality, and Low-Rank Adaptation (LoRA) configurations to optimize model performance and understand their impact on caption quality.

The rest of the paper is organized as follows. Section 2 surveys related work on Arabic image captioning. Section 3 describes the dataset used in this study. Section 4 introduces our methodology, including preprocessing, translation, and fine-tuning steps. Section 5 details the parameter settings, while Section 6 presents results and error analysis. Section 7 discusses ethical considerations, Section 8 outlines limitations, and Section 9 concludes the paper.

2 Related Works

Arabic image captioning is a growing research area focused on creating natural language descriptions for images while respecting the unique features of the Arabic language. This involves dealing with its complex word forms, varied sentence structures, and regional dialects, as well as capturing cultural

Proceedings of The Third Arabic Natural Language Processing Conference, pages 414–418
November 8-9, 2025 ©2025 Association for Computational Linguistics

details to make the captions both accurate and meaningful. The challenges of Arabic captioning stem from its rich morphology, flexible syntax, and diverse dialects. Early advances in image captioning were shaped by the introduction of attention mechanisms (Xu et al., 2015), which have since been adapted for Arabic in multiple studies. A comprehensive review highlighted the need for culturally aware datasets and models tailored to Arabic's linguistic diversity (Al-Khalifa et al., 2021).

Several architectural innovations have been proposed to address these challenges. The *AraCap* framework combined convolutional and recurrent networks to improve fluency and semantic accuracy (Afyouni et al., 2021), while other approaches leveraged visual–textual feature concatenation with pretrained word embeddings for performance gains (Elbedwehy and Medhat, 2023). ResNet50-based visual backbones have also been explored for Arabic captioning tasks (Alazzam, 2022).

Training strategies have evolved alongside architectural improvements. Multi-task learning has been shown to boost caption quality (Za'ter and Talafha, 2022), and self-critical sequence training (SCST) (Rennie et al., 2017) has been adapted for Arabic contexts to refine generation through reinforcement learning. Transfer learning from large-scale vision–language models has further improved performance (Ibrahim et al., 2024), while comparative analyses have examined the impact of deep learning factors on accuracy and robustness (Hejazi and Shaalan, 2021). More recently, BLIP-based vision–language integration has demonstrated strong results for Arabic caption generation (Sayed et al., 2024).

Overall, existing work reflects steady progress in Arabic image captioning, yet also underscores the need for models that are not only computationally efficient but also linguistically and culturally efficient.

3 Dataset Description

We utilized the dataset provided for the Shared Task on Arabic Image Captioning with Cultural Relevance, part of ImageEval 2025 (Bashiti et al., 2025).The dataset contains 3,071 manually annotated images, with 2,717 used for training, 75 for validation and 279 reserved for testing. It includes manually written Arabic captions in the training set that capture the language and cultural details common in Arabic-speaking communities. The test set, provided without captions, ensures a blind evaluation process.Our generated captions were evaluated via the CodaLab platform using standard metrics such as BLEU, cosine similarity to assess their accuracy and cultural relevance.

4 Methodology

4.1 System

Our goal is to produce culturally accurate Arabic captions for historical and cultural images by preserving named entities, and details such as attire and artifacts. Figure 1 provides an overview of the full pipeline. It illustrates how the vision encoder, translation component, and fine-tuning process are linked together. This layout helps clarify how visual information flows into the model and is combined with the translated text before caption generation.

4.2 Image Preprocessing

We have converted all images to grayscale to maintain a consistent visual style and reduce computational complexity by removing non-essential color channels.While RGB inputs can offer richer visual information, our focus was on structural and textural features rather than color-based cues.As, our dataset contained relatively few RGB images, making grayscale a more uniform choice.All images were resized to 224×224 pixels to match the model's input requirements.

4.3 Translation

We translated the original Arabic captions into English using the unsloth/Qwen3-14B model and then back-translated the outputs into Arabic after fine-tuning. This loop improved clarity and fluency, as working in English helped the model generate more precise and descriptive captions while preserving meaning. We chose Qwen3-14B for its strong multilingual training, which made it more effective than alternatives like MarianMT, especially in retaining culturally significant expressions. Although we did not run a large-scale comparison, qualitative checks showed that Qwen3-14B reduced meaning loss during back-translation, which justified its inclusion in the pipeline.

4.4 Image-Caption Pairing

Each preprocessed image was paired with its translated English caption. We then used these image-caption pairs to fine-tune our model, ensuring that the training process learned to generate accurate

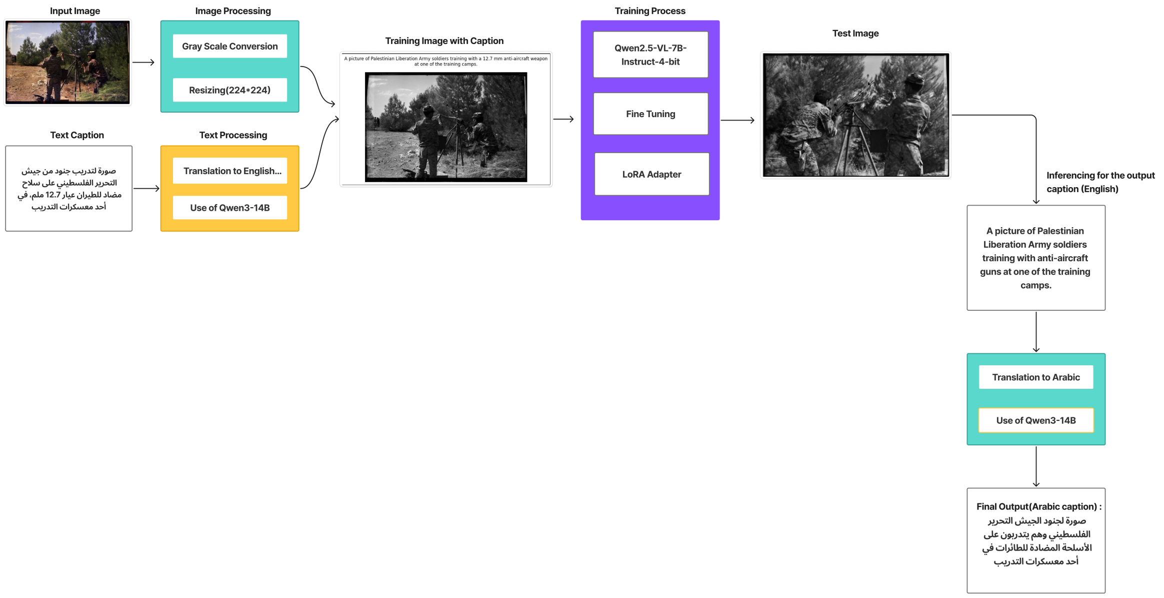
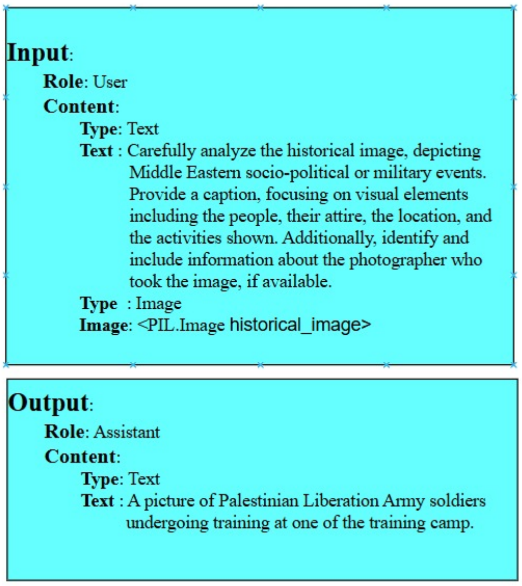

Input Image
Image Processing
Gray Scale Conversion
Resizing(224*224)
Text Caption
صورة لتدريب جنود من جيش
التحرير الفلسطيني على سلاح
مضاد للطيران عيار 12.7 ملم، في
أحد معسكرات التدريب
Text Processing
Translation to English...
Use of Qwen3-14B
Training Image with Caption
A picture of Palestinian Liberation Army soldiers training with a 12.7 mm anti-aircraft weapon at one of the training camps.
Training Process
Qwen2.5-VL-7B-Instruct-4-bit
Fine Tuning
LoRA Adapter
Test Image
Inferencing for the output caption (English)
A picture of Palestinian Liberation Army soldiers training with anti-aircraft guns at one of the training camps.
Translation to Arabic
Use of Qwen3-14B
Final Output(Arabic caption) :
صورة لجنود الجيش التحرير
الفلسطيني وهم يتدربون على
الأسلحة المضادة للطائرات في
أحد معسكرات التدريب

Figure 1: Multimodal architecture for Arabic image captioning using vision-language models



Input:
Role: User
Content:
Type: Text
Text : Carefully analyze the historical image, depicting Middle Eastern socio-political or military events. Provide a caption, focusing on visual elements including the people, their attire, the location, and the activities shown. Additionally, identify and include information about the photographer who took the image, if available.
Type : Image
Image: <PIL.Image historical_image>

Output:
Role: Assistant
Content:
Type: Text
Text : A picture of Palestinian Liberation Army soldiers undergoing training at one of the training camp.

Figure 2: Prompt provided to Qwen2.5-VL-7B-Instruct-bnb-4bit for caption generation

and culturally informed descriptions. Figure 2 illustrates instruction–response formatting for fine-tuning. This formatting ensured consistency between training and inference prompts for caption generation.

4.5 Initial Experimentation

We have evaluated several open-source multimodal and language models for Arabic image caption generation. Specifically,we experimented with unsloth/Llama-3.2-11B-Vision-Instruct and unsloth/Pixtral-12B-2409.Both models were loaded in 4-bit quantized format using the Unsloth library for memory efficiency and configured with gradient checkpointing to support longer context processing.Although these models produced syntactically valid Arabic captions, the outputs lacked semantic adequacy.Pixtral-12B demonstrated strong visual grounding, accurately capturing fine details, but had higher resource demands and slower training. Llama-3.2-11B-Vision converged faster but occasionally omitted culturally specific information BLEU and cosine similarity scores indicated suboptimal performance which motivated us to explore a model with stronger multilingual vision language alignment capabilities.

4.6 Overview of the Adopted Model

We adopted unsloth/Qwen2.5-VL-7B-Instruct-bnb-4bit as our final system due to its efficient multimodal integration, strong instruction-following capabilities, and relatively low computational cost compared to larger models. The model was loaded in 4-bit NF4 quantization using BitsAndBytes. LoRA adapters were applied to both vision and language heads with rank = 64, alpha = 64, and zero dropout, enabling parameter-efficient fine-tuning while keeping the base model largely frozen.

Fine-tuning was performed using the TRL `SFTTrainer` for 3 epochs with a batch size of 32, a learning rate of 5×10^{-5}, a cosine scheduler, and the AdamW 8-bit optimizer with weight decay of 0.01. Gradient checkpointing and FP16 mixed precision were used to reduce memory usage and accelerate training.

5 Parameter Setting

Initial Experiments:

LoRA rank = 16, alpha = 32, dropout = 0.05, batch size = 4, learning rate = 5×10^{-5}, 3 epochs.

Final Model

LoRA rank = 64, alpha = 64, dropout = 0.0, batch size = 32, learning rate = 5×10^{-5}, 3 epochs,

6 Results and Analysis

Table 1 compares the performance of the models on Arabic image captioning. We can see that Qwen 2.5-VL (7B) consistently outperforms both LLaMA 3.2 (11B) and Pixtral (12B). It achieves the highest BLEU-1 score (22.6), the best cosine similarity (57.48), and the highest LLM judge score (31.43), indicating that its captions are not only more accurate but also better aligned with human judgment.

Table 1: Evaluation Metrics for Arabic Image Captioning Models

Model	BLEU-1 Mean	Cosine Sim. Mean	LLM Judge Score
Qwen2.5 VL (7B)	22.6	57.48	31.43
Llama 3.2 (11B)	18.9	49.32	26.75
Pixtral (12B)	15.4	42.19	22.10

LLaMA 3.2 (11B) performs moderately well, but its captions sometimes miss finer cultural or contextual details. Pixtral (12B), struggles to generate semantically accurate and culturally relevant captions. Overall, these results highlight that Qwen 2.5-VL strikes the best balance between understanding the images and producing fluent, culturally aware Arabic captions, making it the most suitable choice for future enhancements in Arabic image captioning systems.

6.1 Error Analysis

Even though Qwen 2.5-VL generates high-quality captions, we noticed some recurring mistakes. Sometimes the model mixes up different Arabic dialects during translation, which can make the captions sound slightly inconsistent. It also occasionally drops culturally important terms. These errors show that while the model understands the images well, capturing the finer linguistic and cultural details in Arabic remains a challenge.

7 Ethical Considerations

For this study, we used the dataset provided in the shared task, which is publicly available. We ensured that all data usage complied with the task guidelines.Since Arabic is culturally and linguistically diverse, we paid special attention to avoid biased or offensive captions.We also recognize that automated captions may occasionally miss cultural or contextual nuances, so we recommend using them to support human judgment rather than replacing it, especially in sensitive contexts.

8 Limitations

Although the proposed pipeline performed well in the shared task, it has several limitations. Reliance on translation and back-translation makes the system dependent on intermediate translator quality, with errors sometimes propagating into final captions. Dialectal variation is also challenging, as the model often defaults to Modern Standard Arabic, limiting its reflection of regional varieties like Palestinian or Levantine Arabic. The dataset (3,071 images) is relatively small compared with large-scale English or Chinese resources, restricting generalization to unseen cultural contexts. Computational constraints also prevented testing larger models or diverse ensembles that could boost performance. Future work should explore larger, culturally diverse datasets, direct Arabic captioning, and improved handling of dialectal diversity.

9 Conclusion

This study has demonstrated the effectiveness of our approach to Arabic image captioning. By fine-tuning Qwen2.5-VL-7B-Instruct and employing translation-based training strategies, we achieved strong performance across multiple evaluation metrics. Integrating cultural preservation techniques and efficient fine-tuning proved essential for capturing the subtle linguistic details in Arabic captions.

Our comparison with other vision-language models highlights the clear advantage of instruction-tuned large language models in generating fluent, context-aware, and culturally sensitive descriptions. At the same time, challenges such as translation dependency and dialectal variations remain, pointing to opportunities for future work.

References

Imad Afyouni, Imtinan Azhar, and Ashraf Elnagar. 2021. Aracap: A hybrid deep learning architecture for arabic image captioning. *Procedia Computer Science*, 189:171–178.

Hend Al-Khalifa, Mustafa Jarrar, and Wajdi Zaghouani. 2021. Challenges in arabic image captioning: A review. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–25.

Batool Mohammed Alazzam. 2022. Arabic image captioning using resnet50. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 14(1):81–88.

Ahlam Bashiti, Alaa Aljabari, Hadi Hamoud, Md. Rafiul Biswas, Bilal Shalash, Mustafa Jarrar, Fadi Zaraket, George Mikros, Ehsaneddin Asgari, and Wajdi Zaghouani. 2025. Imageeval 2025: The first arabic image captioning shared task. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Samar Elbedwehy and Tamer Medhat. 2023. Improved arabic image captioning model using feature concatenation with pre-trained word embedding. *Neural Computing and Applications*, 35(25):18575–18592.

Hani D. Hejazi and Khaled Shaalan. 2021. Deep learning for arabic image captioning: A comparative study of main factors and preprocessing recommendations. *International Journal of Advanced Computer Science and Applications*, 12(11):37–44.

Haneen Siraj Ibrahim, Narjis Mezaal Shatia, and AbdulRahman A. Alsewari. 2024. A transfer learning approach for arabic image captions. *Mustansiriyah Journal of Science*, 35(1):70–79.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA. IEEE.

Abdelrahman M. Sayed, Mohamed K. Elhadad, Gouda I. Salama, and Aiman M. Mousa. 2024. Improving arabic image captioning with vision-language models. In *Proceedings of the 10th International Conference on Electrical Engineering and Informatics (ICEEI)*, Bandung, Indonesia. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France. PMLR.

Muhy Eddin Za’ter and Bashar Talafha. 2022. Benchmarking and improving arabic automatic image captioning through the use of multi-task learning paradigm. *arXiv preprint arXiv:2202.05474*.