Codezone Research Group at ImageEval Shared Task: Arabic Image Captioning Using BLIP and M2M100 A Two-Stage Translation Approach for ImageEval 2025

Abdulkadir Shehu Bichi[1], **Ismail Dauda Abubakar**[2], **Fatima Muhammad Adam**[3], **Aminu Musa**[3], **Auwal Umar Ahmed**[4], **Abubakar Ibrahim**[4], **Khadija Salihu Auta**[5], **Aisha Mustapha Ahmed**[6], **Mahmud Said Ahmed**[7]

[1]**Baba Ahmed University Kano,** [2]**Federal University Gusau,** [3]**Federal University Dutse,**

[4]**Northbridge College of Science and Technology,** [5]**Khalifa Isyaku Rabiu University, Kano (KHAIRUN),**

[6]**Bayero University Kano,** [7]**Federal University of Technology - Babura**

Correspondence: abdulkadir.bichi@babaahmeduniversity.edu.ng

Abstract

This paper details the ImageEval 2025 Shared Task on Arabic image captioning. We designed a two-step, zero-shot framework that utilises the BLIP multimodal vision-language model to first generate English captions. These captions are then converted to Arabic via the M2M100 multilingual translation model. We tested the full pipeline on the official ImageEval 2025 benchmarking set, obtaining a cosine similarity of 0.383 and an LLM Judge score of 15.14. The corroborating numerical and qualitative findings confirm the viability of a translation-driven methodology for cross-lingual image captioning in Arabic, a language often classified as low-resource. Nonetheless, the experiments also uncovered weaknesses: subtle semantic layers and culturally specific references are inadequately conveyed in the output and merit focused attention in subsequent iterations.

Keywords: Arabic image captioning, captioning algorithms, BLIP, M2M100, cross-lingual transfer, multilingual machine translation

1 Introduction

The task of image captioning presents a significant challenge in the fields of computer vision and natural language processing, where models are expected to describe images using natural language. Although considerable effort has been devoted to English image captioning, generating Arabic captions that are both culturally relevant and contextually accurate remains a major challenge due to limited resources and the unique characteristics of the Arabic language (Bashiti et al., 2025).

The ImageEval 2025 Shared Task initiates an Arabic image captioning evaluation framework by providing a dataset of 3,471 images paired with Arabic captions (Bashiti et al., 2025). This initiative aims to facilitate the development of Arabic vision-language models capable of generating culturally relevant and linguistically accurate textual descriptions of images.

Addressing the outlined problem is achievable through a two-step captioning strategy: first, using the leading BLIP model to generate English descriptions of the images, which are then translated into Arabic using the M2M100 multilingual machine translation model. This approach takes advantage of the extensive ImageEval 2025 Shared Task datasets and the strong translation capabilities for the generation of Arabic text. Therefore, our review of the literature suggests that this method is a new technique for Arabic image captioning, since no previous research has used this method to generate captions in Arabic from images.

The remainder of this paper is organized as follows: Section 2 reviews related work, while Section 3 presents our two-stage BLIP+M2M100 methodology. Section 4 describes the experimental setup, followed by Section 5 which presents quantitative and qualitative results with comparative analysis against baseline models. Finally, Section 6 concludes with key findings and discusses future research directions for Arabic image captioning.

2 Related Work

The recent development of vision-language models has predominantly focused on the English language, utilizing models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP (Li et al., 2022). These models achieve zero-shot performance by learning joint representations of images and text from large-scale web datasets. However, their application to Arabic remains almost nonexistent, primarily due to the scarcity of available resources and data.

Research on multilingual image captioning has primarily focused on multilingual training (Li et al., 2020), cross-lingual transfer learning (Stefanini et al., 2023), and other translation-based methods (Elliott and de Vries, 2023). Although translation-based approaches to image captioning are straight-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 390–394

forward, they are effective when high-quality translation software is available. A significant advancement in multilingual neural machine translation is the M2M100 model (Fan et al., 2021), which can translate directly between one hundred languages without relying on English as a central pivot language.

The processing of Arabic text presents challenges such as complex morphology, diverse dialects, and the right-to-left writing direction (Habash, 2010). These characteristics complicate Arabic text generation and evaluation, underscoring the importance of advancing image-captioning systems in Arabic. This, in turn, fosters the development of Arabic vision-language understanding systems.

3 Methodology

This section explains the system architecture, the process of caption creation for images, translation from English to Arabic, and text normalization.

3.1 System Architecture

The system architecture consists of the following two components.

1. English Caption Generation: The BLIP model generates English descriptions for the given images.

2. Arabic Translation: English captions are translated into Arabic using the M2M100 model. As shown in Figure 1, our proposed system employs a two-stage pipeline approach.

The modular approach enables the utilization of existing English vision-language models alongside state-of-the-art neural machine translation techniques for generating Arabic text.

3.2 Image Captioning and English-to-Arabic Translation

For English captioning, we use the BLIP-base model (Salesforce/blip-image-captioning-base). BLIP employs a unified vision-language pretraining strategy that integrates the training of an image encoder, a text encoder, and an image-grounded text decoder. It is pretrained on large image-text corpora, enabling the model to generate accurate captions for images without prior exposure to specific content. For each provided image, we perform the following steps:

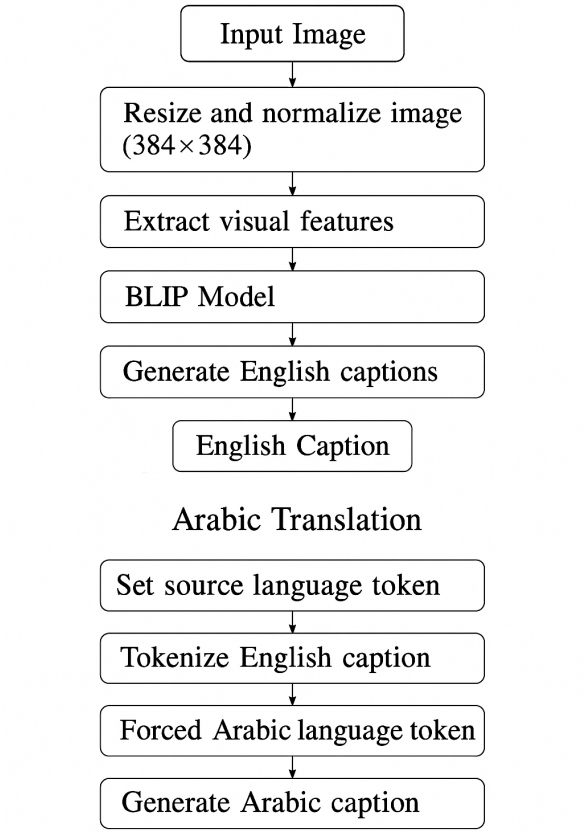

Figure 1: Architecture of the Proposed Model: Arabic Image Captioning Using BLIP and M2M100

1. The image is resized to a resolution of 384 × 384 pixels and then normalized as employed by (Rastogi, 2024).

2. The BLIP vision encoder extracts the corresponding visual features from the image.

3. English captions are generated using beam search decoding.

4. Translation of the caption is performed on the selected caption that has received the highest score.

5. The source language token is set to English ("en").

6. Tokenization of the English caption using the M2M100 tokenizer.

7. Generation of Arabic translations with the mandatory use of the Arabic language token.

8. Encoding the result to generate the Arabic caption.

3.3 Text Normalization

For evaluation purposes, we performed Arabic text normalization as proposed by (Alami Chehbouni et al., 2020), which includes the following steps: Removing diacritical marks, Removal of Tatweel characters, Removal of punctuation marks, Whitespace Standardization.
This normalization accounts for the morphological intricacies of the Arabic language, providing a robust and fair evaluation against the reference captions.

4 Experimental Setup

This section describes the dataset used, implementation details, and evaluation metrics.

4.1 Dataset

We conducted a system evaluation using the ImageEval 2025 dataset, which comprises 3,471 images with captions, distributed as follows (Bashiti et al., 2025).
1. Training set: 2,718 images
2. Validation set: 75 images
3. Test set: 752 images
The dataset includes a diverse array of visuals accompanied by culturally relevant Arabic captions, making it a robust benchmark for evaluating Arabic image captioning.

4.2 Implementation Details

The following system configuration was used for the implementation:

1. Hardware: Image processing with CUDA-enabled GPUs.
2. BLIP Model: Salesforce/BLIP Image Captioning Based
3. Translation Model: facebook/m2m100-418M
4. Framework: PyTorch with the Transformers library.
5. Inference: Conducted in a zero-shot scenario without any prior model tuning.

The entire processing pipeline generates captions for 752 test images, with an average processing time of 12.27 seconds per image for both caption generation and translation.

4.3 Evaluation Metrics

We used multiple metrics to evaluate the quality of the captions.

1. BLEU Scores: N-gram precision metrics (BLEU-1 through BLEU-4).
2. ROUGE Scores: Recall-oriented metrics including ROUGE-1, ROUGE-2, and ROUGE-L.
3. Cosine Similarity: A metric for evaluating multi-lingual sentence embeddings.
4. LLM Judge Score: Evaluation conducted by large language models.

These metrics analyze various aspects of a caption, including its text, meaning, and human evaluation.

5 Results and Analysis

This section explains qualitative results, error analysis, and provides qualitative examples.

5.1 Quantitative Results

Table 1 reports the metrics achieved by our system in the test set, together with comparisons with the baseline models.

Our dual-pass translation framework achieves significant improvements over standard models, as measured by classical n-gram metrics such as BLEU and ROUGE. In particular, the system achieves substantial gains in BLEU-1 (0.2847 compared to 0.0992 for zero shot and 0.1698 for fine tuned Qwen 2.5-VL), outperforming both zero shot and fine tuned Qwen 2.5-VL models. However, the baseline variants exhibit higher cosine similarity and LLM judge scores, highlighting a complementary balance between precise semantic representation and holistic quality assessment offered by the two architectures.

Table 1: Quantitative Results Comparison – Our Arabic Image Captioning Method vs. Baseline Models

Metric	Our Method (BLIP + M2M100)	Zero-shot Qwen 2.5-VL 7B (Baseline)	Fine-tuned Qwen 2.5-VL 7B (Baseline)
BLEU-1	0.2847	0.0992	0.1698
BLEU-2	0.1623	0.0323	0.0862
BLEU-3	0.0943	0.0190	0.0543
BLEU-4	0.0587	0.0133	0.0305
ROUGE-1	0.0000	0.0000	0.0000
ROUGE-2	0.0000	0.0000	0.0000
ROUGE-L	0.0000	0.0000	0.0000
Cosine Similarity	0.3830	0.5577	0.5846
LLM Judge Score	15.1400	27.1100	30.8200

5.2 Qualitative Assessment Outcomes

Beyond numerical evaluation, the framework was assessed using qualitative criteria that addressed both the cultural and linguistic appropriateness of the generated captions.

1. Cultural Relevance: 1.10
2. Conciseness: 2.03
3. Completeness: 1.47
4. Accuracy: 2.03

The modest scores highlight a pressing need to improve cultural depth and caption comprehensiveness, while conciseness and accuracy demonstrate steady, if not outstanding, performance.

5.3 Error Analysis

This section, examining the generated captions, uncovers some recurring issues.

1. Idioms: Some English phrases and sayings do not have an equivalent in Arabic.
2. Cultural Relevance: Some generated captions include references that lack culturally relevant details or specific information. Morphological variations in Arabic pose challenges to exact lexical matching due to its complex morphology.
3. Object Misrecognition: Some devices or ideas that belong to specific cultures are misrecognized.

5.4 Comparative Analysis

Our two-stage approach demonstrates distinct performance characteristics compared to the baseline models.

Strengths:

1. Led performance evaluation using surface n-gram overlap metrics (BLEU, ROUGE).
2. Leveraging advanced English vision-language encoding techniques
3. Stable pipeline extending from Arabic vision to generated captions

Limitations:

1. Achieved through dedicated multilingual captioning models.
2. Performance on large language model evaluation metrics continues to lag behind state-of-the-art benchmarks.
3. Qualitative assessments identify instance-specific gaps in cultural relevance, reaffirming the necessity of localized context.

However, results indicate that translation-based pipelines produce captions with high linguistic fidelity to reference standards, whereas models that retain multilingual embeddings convey deeper semantic information, albeit with slightly lower lexical precision.

6 Limitation

No Direct Visual-Arabic Learning: The approach cannot learn how Arabic speakers naturally describe visual content, relying instead on English visual understanding followed by translation, which misses Arabic-specific visual-linguistic patterns.

7 Conclusion and Future Work

The shift and contribution of this paper lies within the use of the BLIP and M2M100 to produce a new two-stage Arabic image captioning system as well as the comparative Qwen 2.5-VL (zero-shot and fine-tuned) baseline analysis of the Arabic datasets. The regional context has not previously been analyzed.

Furthermore, this study presents a two-phase Arabic captioning architecture that takes advantage of existing English vision-language models alongside tile-based multilingual translation services. The resulting system achieved a competitive cosine similarity score of 0.383, demonstrating the feasibility of translation-centric cross-lingual image captioning. Performance metrics indicate that translation-based approaches outperform conventional n-gram baselines; however, evaluations of semantic coherence and cultural representation reveal gaps that require targeted refinement. Future iterations will incorporate deeper multilingual embeddings and culturally aware context modules to enhance both meaning preservation and cultural resonance.

Future enhancements may include the following:

1. Direct Arabic Vision-Language Models: Develop fully end-to-end Arabic image captioning systems that utilize large-scale, culturally specific datasets to maximize relevance.
2. Cultural Context Enhancement: Integrate structured cultural knowledge graphs with annotation pipelines to ensure that relevance scoring and caption generation reflect nuanced local traditions.
3. Hybrid Approaches: Combine the rigor of lexically precise, translation-inspired systems with the deep semantic capabilities of multilingual transformers within a balanced,

modular, and selective architecture.

4. Advanced Text Normalization: Implement state-of-the-art morphological disambiguation and dialect-aware normalization techniques to standardize Arabic text while minimizing contextual distortion.

The ongoing research highlights the rapid progress in Arabic artificial intelligence. The upcoming ImageEval 2025 benchmark is expected to further intensify competition in Arabic vision and language comprehension.

Acknowledgments

The ImageEval 2025 organization has provided a detailed evaluation and dataset that have contributed significantly to the advancement of Arabic artificial intelligence. The creators of the dataset and evaluation framework have offered invaluable resources for Arabic Natural Language Processing.

Code Availability Statement

The code supporting the findings of this study is openly available at https://github.com/asbichi362/Arabic-Image-Captioning-Using-BLIP-and-M2M100

References

A. Alami Chehbouni, A. Ouatik Said, and T. Rachidi. 2020. A proposed natural language processing preprocessing procedures for enhancing arabic text summarization. In *Advances in Intelligent Systems and Computing*, pages 25–34. Springer.

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.

J. Elliott and A. P. de Vries. 2023. Evaluating image captioning systems via cross-modal retrieval. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12543–12559.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

N. Y. Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916.

J. Li, D. Li, C. Xiong, and S. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900.

X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–137.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, G. Krueger J. Clark, and I. Sutskever. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Ritvik Rastogi. 2024. Papers explained 190: Blip-3 (xgen-mm). https://ritvik19.medium.com/papers-explained-190-blip-3-xgen-mm-6a9c04a3892d. Medium.

M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. 2023. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):539–559.