Averroes at ImageEval 2025 Shared Task: Advancing Arabic Image Captioning with Augmentation and Two-Stage Generation

Mariam Saeed[1,2], Sarah Elshabrawy[3], Abdelrahman Hagrass[1,3], Mazen Yasser[1,2], Ayman Khalafallah[1,2]

[1]Applied Innovation Center, [2]Alexandria University [3]Georgia Institute of Technology

Applied Innovation Center [m.saeed,a.hagrass,m.yasser,a.khalafallah]@aic.gov.eg

Alexandria University: [es-mariamzaho4,es-mazen2215,ayman.khalafallah]@alexu.edu.eg

Georgia Tech: [selshabrawy3,ahagrass3]@gatech.edu

Abstract

Image captioning aims to generate natural language descriptions of images, combining visual understanding with language generation. This task is particularly challenging in low-resource settings such as Arabic, where annotated data is limited and captions must reflect both cultural and linguistic nuances. In this system paper, we present our approach for the ImageEval 2025 Arabic Image Captioning Shared Task. Our system is based on the Qwen2.5-VL-7B vision-language model, enhanced with quality-aware data augmentation, a two-stage description-to-caption pipeline, and post-processing for improved fluency. In the official evaluation, our approach ranked **first** in the *LLM as a Judge* metric with a score of **33.97**, **second** in *Cosine Similarity* with a score of **58.55**, and **first** in the manual evaluation phase conducted by the organizers.

1 Introduction

Image captioning generates natural language descriptions of images by combining visual understanding with language generation. While vision-language models (VLMs) have achieved strong results in high-resource languages, applying them to Arabic remains challenging due to limited annotated data, complex morphology, and the need for culturally appropriate captions.

The ImageEval 2025 Arabic Image Captioning Shared Task (Bashiti et al., 2025) addressed these challenges by releasing a manually annotated Arabic captioning dataset and a standardized evaluation framework. Systems were evaluated using BLEU (Papineni et al., 2002), Cosine Similarity, and LLM-as-a-Judge scores (Li et al., 2024) during the submission phase, followed by a manual evaluation by the organizers.

We present our system for this task, built on the Qwen2.5-VL-7B model (Team, 2025) with quality-aware data augmentation, a two-stage description-to-caption pipeline, and regex-based post-processing. We also explored lighter models such as BLIP, but Qwen2.5-VL-7B proved superior. Our system ranked **first** in LLM-as-a-Judge (33.97), **second** in Cosine Similarity (58.55), and **first** in manual evaluation, demonstrating the effectiveness of combining large VLMs with targeted augmentation and structured generation for Arabic captioning.

The rest of the paper is organized as follows: Section 3 details our system, Section 4 presents the dataset, metrics, and results, and Section 6 concludes.

2 Related Work

Image captioning aims to produce natural language descriptions of images by combining visual recognition with language generation. Early approaches paired CNN-based encoders with RNN decoders (Vinyals et al., 2015; Karpathy and Fei-Fei, 2017), later enhanced by attention mechanisms (Xu et al., 2015; Anderson et al., 2018) and, more recently, transformer architectures (Cornia et al., 2020).

The field has since shifted toward large vision-language models (VLMs) that integrate powerful image encoders with pretrained language models, enabling stronger cross-modal reasoning. Prominent examples include CLIP (Radford et al., 2021), BLIP (Li et al., 2022), Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023), and Qwen-VL (Bai et al., 2023; Team, 2025), which leverage large-scale multimodal pretraining and instruction tuning to achieve state-of-the-art performance.

Due to their size, adapting VLMs for specific tasks often relies on parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022), implemented in frameworks like PEFT (Mangrulkar et al., 2022), which significantly reduce computational and memory requirements while preserving model quality.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 432–437

Evaluating image captioning systems has traditionally relied on automatic metrics such as BLEU (Papineni et al., 2002), which measure n-gram overlap, and cosine similarity with TF–IDF (Sparck Jones, 1988; Salton and Buckley, 1988), which captures content similarity beyond surface form. More recently, human-aligned evaluation methods such as LLM-as-a-Judge (Li et al., 2024) have gained attention, assessing captions on semantic accuracy, fluency, and cultural relevance in a manner closer to human judgment.

3 System Overview

In this section, we outline the progression of our experiments during the shared task, starting from initial baselines and gradually introducing more advanced augmentation strategies and post-processing techniques. While we kept the underlying model architecture unchanged, our approach evolved from a single-model setup to a two-model pipeline for improved performance.

3.1 Baseline: Single-Stage Captioning

We began by fine-tuning **Qwen2.5-VL-7B** [1] using LoRA to assess its ability to generate Arabic image captions without any additional enhancements. LoRA allowed us to update only a small subset of parameters while keeping most of the model frozen, reducing computational cost while adapting it to the task dataset. The baseline training prompt was intentionally simple:

Baseline Prompt

Describe the image in Arabic.

The organizers released scores for both a fully fine-tuned Qwen model and a zero-shot baseline. Our LoRA-based variant yielded different outcomes, which we detail in the results section, and served as the reference point for all subsequent enhancements.

3.2 Smaller Architectures

We wanted to explore the feasibility of using smaller vision-language models for the task, so we experimented with BLIP (Li et al., 2022). We started from a checkpoint already fine-tuned on Flickr8k Arabic captioning dataset[2] and further fine-tuned it on the task dataset. Although BLIP converged quickly, its performance, particularly in capturing fine-grained Arabic details, was noticeably worse than Qwen2.5-VL-7B. Based on these results, we decided to focus on Qwen2.5-VL-7B for the subsequent experiments.

3.3 Data Augmentation Strategies

Given the limited size of the training dataset, we employed two augmentation strategies to improve generalization and assess the performance of different training pipelines.

Aug1: Classical Image Augmentation. The first approach, **Aug1**, applied three random transformations to each image from a predefined set implemented in the `Albumentations` library. The transformations included cropping or padding, horizontal flipping, rotation, small-scale shifting and zooming, motion blur, and Gaussian noise. Captions were kept unchanged, tripling the dataset size and exposing the model to more varied visual patterns while preserving semantic content.

Aug2: Quality-Aware Caption and Image Augmentation. While Aug1 increased visual diversity, it did not introduce textual variation. In **Aug2**, we first augmented captions: for each image, we used **Aya-Vision-8B**[3] to generate three slightly different captions and computed their BLEU score against the original. Captions scoring below 0.75 were discarded to ensure semantic consistency. For each retained augmented caption, one random Aug1 transformation was applied to its image. This process added 814 high-quality samples to the training set. Later, we combined these augmentation strategies with different training pipelines.

3.4 Structured Caption Generation with Descriptions

We hypothesized that guiding the model to first produce a detailed description of the image would lead to more accurate captions. To train such a system, we first created a dataset of image–description–caption triples using **Aya-Vision-8B**. The descriptions were generated with the following prompt:

Description Data Prompt

Describe this image in detail in Arabic to help in extracting the following caption between <cap> tags.
<cap>
caption
</cap>

[1] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

[2] https://huggingface.co/omarsabri8756/blip-Arabic-flickr-8k

[3] https://huggingface.co/CohereLabs/aya-vision-8b

This prompt was designed to produce not only a general description, but also to highlight the key details and important elements that would support accurate caption generation.

We then fine-tuned **Qwen2.5-VL-7B** using a structured output format that explicitly separated the description from the caption:

Structured Training Prompt

Describe the image in detail in Arabic. Then, based on that description, generate a suitable caption in Arabic (10-30 words).
<description>
Detailed description
</description>
<cap>
Caption
</cap>

This structured approach encouraged the model to first organize its observations and reasoning before producing the final caption.

3.5 Two-Model Pipeline

Building on the structured captioning idea, we developed a two-model pipeline. The first model (*Description Model*) generated a detailed description from the image, while the second model (*Caption Model*) used both the image and the description to produce the final caption. Both models were based on Qwen2.5-VL-7B and trained independently.

3.6 Post-Processing and Model Merging

During evaluation, we found that some generated captions contained repetitions or redundant phrases. We applied a regex-based cleaning step to remove such artifacts, improving fluency and readability.

We also observed that two variants of the two-model pipeline excelled in different aspects of captioning, specifically, the pipeline trained with Aug2 and the one without augmentation. To combine their strengths, we performed model merging, a technique that integrates parameters from multiple trained models into a single model, aiming to retain beneficial knowledge from each. We used **MergeKit** (Goddard et al., 2024) with the **TIES** algorithm (Yadav et al., 2023) to merge the models at the parameter level, preserving their complementary capabilities.

3.7 Final System

Our final submission integrated the most effective components from our experiments. It used the **Aug2** quality-aware augmentation to enrich both visual and textual diversity, followed a two-model Qwen2.5-VL-7B pipeline for structured generation (Description → Caption), and applied regex-based cleaning to improve output fluency. The complete workflow is illustrated in Figure 1.

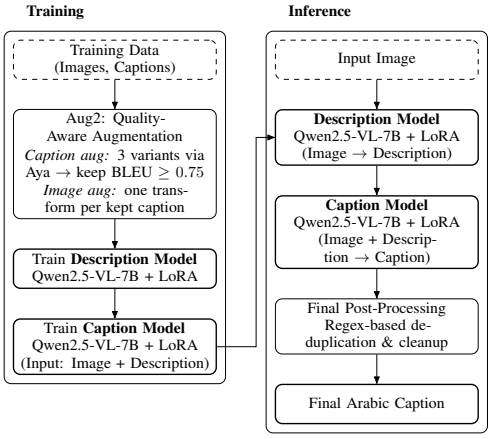

Figure 1: Final system: Aug2 data preparation, two-model Qwen2.5-VL pipeline (Description then Caption), and regex-based cleanup.

4 Experiments

4.1 Dataset and Metrics

We used the training data provided by the organizers of the shared task (Bashiti et al., 2025), a manually captioned dataset of 3,471 images, split into 2,718 for training and 753 for testing. To validate and analyze our approaches during development, we further divided the training set into a 90/10 split, using the smaller portion as a validation set.

Submissions were evaluated using four metrics. During the submission period, **BLEU**, **Cosine Similarity**, and **LLM as a Judge** scores were reported on the public leaderboard. BLEU measured n-gram overlap between the generated caption and the reference, capturing surface-level similarity in wording. Cosine similarity measured the textual closeness between generated captions and reference descriptions after Arabic-specific normalization and TF–IDF vectorization. The LLM-as-a-Judge metric used `gpt-4o` with a fixed seed and zero temperature to score captions on semantic accuracy, relevance, and fluency, with results normalized to a 0–100 scale.

After the submission period, the organizers conducted a **Manual Evaluation** on about 5% of the test set, assessing Cultural Relevance, Conciseness, Completeness, and Accuracy, each on a 1–4 scale.

4.2 Training Setup

All experiments were conducted on a single NVIDIA A100 GPU with 80 GB of memory. The Qwen models were fine-tuned using LoRA with rank 8, targeting all modules. Training was performed with a batch size of 2 and gradient accumulation over 8 steps, giving an effective batch size of 16. For the BLIP model, a batch size of 8 was used. We set the learning rate to 2×10^{-5} with a cosine scheduler and a warmup ratio of 0.1, using `bf16` precision. The input cutoff length was fixed at 2048 tokens. Models were trained for a maximum of 10 epochs, and the checkpoint achieving the lowest loss on our validation set was selected for submission.

4.3 Results and Analysis

Table 1 presents the results of the different variants of our system across BLEU, Cosine Similarity, and LLM-as-a-Judge. The baselines provided by the organizers include a zero-shot Qwen2.5-VL-7B and a fully fine-tuned version. Our LoRA baseline already surpassed both organizer-provided baselines, achieving 22.84 BLEU and 30.19 LLM-as-a-Judge.

Classical image augmentation (**Aug1**) applied to the LoRA baseline slightly reduced LLM-as-a-Judge and Cosine Similarity scores, suggesting that random visual perturbations without textual augmentation do not consistently help. Applying Aug1 to BLIP yielded lower scores overall, confirming that BLIP was less competitive for this task.

Structured output improved semantic evaluation, with the structured-only variant achieving 32.96 LLM-as-a-Judge. Adding quality-aware augmentation (Aug2) increased BLEU to 23.76 but slightly reduced LLM-as-a-Judge, indicating a trade-off between n-gram overlap and semantic quality.

The **two-model pipeline** proved particularly effective, achieving the highest BLEU (24.99) among our systems without augmentation and 33.81 LLM-as-a-Judge when combined with **Aug2**. Merging two-model pipelines trained with and without Aug2 preserved strong BLEU and Cosine scores but slightly lowered LLM-as-a-Judge.

Our final system, two-model pipeline with Aug2 and regex-based output cleaning, achieved the highest LLM-as-a-Judge score (33.97), second place in Cosine Similarity (58.55), and competitive BLEU (24.39), confirming the benefit of structured generation, quality-aware augmentation, and light post-processing.

Model	BLEU	Cosine	LLM-as-a-Judge
Baseline zero-shot (organizers)	9.92	55.77	27.11
Baseline full (organizers)	16.89	58.46	30.82
Baseline LoRA	22.84	56.95	30.19
Baseline + Aug1	22.50	56.33	28.58
BLIP + Aug1	19.95	54.42	19.83
Structured output + Aug2	23.76	57.33	31.71
Structured output	23.31	58.23	32.96
Two-model pipeline	**24.99**	57.72	32.25
Two-model pipeline + Aug2	24.22	**58.58**	33.81
Merging two-model variants	24.55	58.53	33.11
Final system	24.39	58.55	**33.97**

Table 1: Performance of different approaches on the shared task test set.

4.4 Qualitative Analysis

While automatic metrics such as BLEU provide a numerical indication of similarity, qualitative inspection offers a clearer view of how the model succeeds and where it fails. We selected four representative samples from the validation set. The first row of Figure 2 presents two failure cases: in the first, the prediction is entirely off-topic, describing a political exhibition instead of sculptures; in the second, the model confuses a youth music rehearsal with a militant training scene.

The second row shows two successful cases. In the first, the model correctly identifies a military training camp, differing only slightly in phrasing, which still yields a high BLEU score. In the second, the prediction matches the ground truth exactly for a school celebration, resulting in a perfect score. These examples indicate that the model performs reliably on visually distinctive and well-defined scenes but struggles when different events share overlapping contextual cues.

4.5 Official Leaderboard Results

Table 2 shows the official Subtask 2 leaderboard for Cosine Similarity and LLM-as-a-Judge. Our team (**Averroes**) ranked **first** in LLM-as-a-Judge with a score of 33.97 and **second** in Cosine Similarity (58.55).

In the manual evaluation phase (Table 3), our system obtained the highest scores in Cultural Relevance (3.63) and Conciseness (3.43), securing **first place** in this stage of the competition.

5 Limitations

While our approach demonstrates promising results, it is constrained by the limited size of the dataset (2,718 samples), which reduces the model's ability to capture diverse semantic and visual patterns. The augmentation strategies we introduced

Failure 1	
GT (AR)	منحوتات تجسد القرآن والعهد القديم
GT (EN)	Sculptures depicting the Qur'an and the Old Testament
Pred (AR)	صورة للجناح الفلسطيني في معرض دمشق الدولي عام ١٩٧١، برفقة شخصيات من منظمة التحرير الفلسطينية.
Pred (EN)	An image of the Palestinian wing at the Damascus International Fair in 1971, accompanied by figures from the Palestine Liberation Organization. (BLEU = 0.0)

Failure 2	
GT (AR)	صورة يظهر فيها تدريب فريق موسيقي من الشباب الفلسطيني استعدادًا للمشاركة في مهرجان الشباب العربي في الجزائر عام ١٩٧٢
GT (EN)	An image showing the training of a Palestinian youth music group in preparation for participating in the Arab Youth Festival in Algeria in 1972
Pred (AR)	صورة لأعضاء من الجبهة الشعبية للتحرير الفلسطيني، داخل أحد مراكز التدريب في بيروت.
Pred (EN)	An image of members of the Popular Front for the Liberation of Palestine inside a training center in Beirut. (BLEU = 0.2094)

Success 1	
GT (AR)	صورة لتدريب جنود جيش التحرير الفلسطيني في أحد معسكرات التدريب
GT (EN)	An image of the training of Palestinian Liberation Army soldiers in one of the training camps
Pred (AR)	صورة لجنود جيش التحرير الفلسطيني في أحد معسكرات التدريب
Pred (EN)	An image of Palestinian Liberation Army soldiers in one of the training camps. (BLEU = 0.7954)

Success 2	
GT (AR)	صورة لحفل مدرسي بمناسبة فلسطينية في إحدى مدارس الفتيات بالكويت.
GT (EN)	An image of a school celebration on a Palestinian occasion in a girls' school in Kuwait.
Pred (AR)	صورة لحفل مدرسي بمناسبة فلسطينية في إحدى مدارس الفتيات بالكويت.
Pred (EN)	An image of a school celebration on a Palestinian occasion in a girls' school in Kuwait. (BLEU = 1.0)

Figure 2: Qualitative examples of Arabic–English captioning. Top row: failure cases with low BLEU scores, where predicted captions diverge from the ground truth. Bottom row: successful cases with high BLEU scores and strong semantic alignment.

Team	Cosine Similarity	LLM-as-a-Judge
VLCAP	**60.01**	33.05
Averroes (ours)	58.55	**33.97**
Phantom Troupe	57.48	31.43
ImpactAi	56.22	26.55
Codezone Research Group	38.30	15.14

Table 2: Official Subtask 2 leaderboard for Cosine Similarity and LLM-as-a-Judge.

Team	Cultural Relevance	Conciseness	Completeness	Accuracy
Averroes (ours)	**3.63**	**3.43**	2.60	2.80
Phantom Troupe	3.40	3.27	2.33	2.40
VLCAP	2.57	3.17	**2.67**	**2.97**
Codezone Research Group	1.10	2.03	1.47	2.03
ImpactAi	3.13	2.73	1.77	1.97

Table 3: Manual evaluation scores on 5% of the test set (1=lowest, 4=highest).

mitigate this limitation to some extent, but cannot fully substitute for a larger, more representative dataset.

Another limitation lies in the reliance on synthetic captions. Although we applied quality control to ensure semantic consistency, automatically generated captions may still introduce noise or overlook subtle aspects of the images.

Finally, our experiments were conducted with a single model size (Qwen2.5-7B). The effect of scaling the model or exploring alternative architectures on caption quality remains an open question for future work.

6 Conclusion

We presented our Qwen2.5-VL-7B–based system for the ImageEval 2025 Arabic Image Captioning Shared Task, integrating quality-aware augmentation, a two-stage description-to-caption pipeline, and regex-based post-processing. The system ranked **first** in *LLM-as-a-Judge*, **second** in *Cosine Similarity*, and **first** in manual evaluation, highlighting the effectiveness of combining large vision-language models with targeted augmentation and structured generation. Future work will explore scaling to larger datasets, multilingual pretraining, and RLHF for improved human alignment.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual

language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Karen Sparck Jones. 1988. *A statistical interpretation of term specificity and its application in retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Qwen Team. 2025. Qwen2.5-vl.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Preprint*, arXiv:2306.01708.