ImageEval 2025: The First Arabic Image Captioning Shared Task

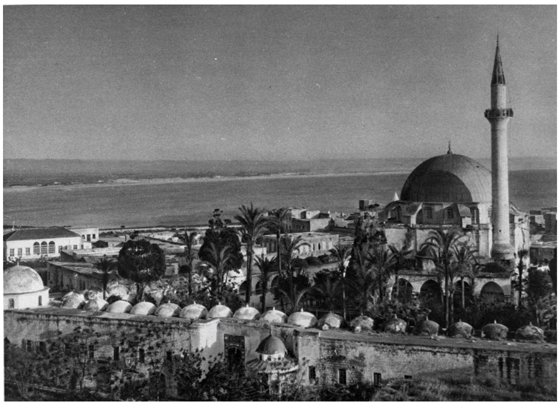
Ahlam Bashiti[1], Alaa Aljabari[1], Hadi Hamoud[3], Md. Rafiul Biswas[2] , Bilal Shalash[3], Mustafa Jarrar[2,1], Fadi Zaraket[3,4], George Mikros[2], Ehsaneddin Asgari[2], Wajdi Zaghouani[5]

[1]Birzeit University, [2]Hamad Bin Khalifa University, [3]American University of Beirut,

[4]Arab Center for Research and Policy Studies, [5]Northwestern University in Qatar

Abstract

We present ImageEval 2025, the first shared task dedicated to Arabic image captioning. The task addresses the critical gap in multimodal Arabic NLP by focusing on two complementary subtasks: (1) creating the first open-source, manually-captioned Arabic image dataset through a collaborative datathon, and (2) developing and evaluating Arabic image captioning models. A total of 44 teams registered, of which eight submitted during the test phase, producing 111 valid submissions. Evaluation was conducted using automatic metrics, LLM-based judgment, and human assessment. In Subtask 1, the best-performing system achieved a cosine similarity of 65.5, while in Subtask 2, the top score was 60.0. Although these results show encouraging progress, they also confirm that Arabic image captioning remains a challenging task, particularly due to cultural grounding requirements, morphological richness, and dialectal variation. All datasets, baseline models, and evaluation tools are released publicly to support future research in Arabic multimodal NLP.



Manual Caption (Culturally Relevant):
صورة تظهر مسجد الجزّار في مدينة عكا الساحلية في فلسطين، أحد أبرز المعالم العثمانية بقبته ومئذنته البارزتين، تحيط به بيوت وأسوار قديمة، ما يعكس الطابع الحضاري والتاريخي للمدينة.
Translation: Al-Jazzar Mosque in Acre, Palestine, a major Ottoman landmark with its dome and minaret, surrounded by old houses and city walls reflecting the city's history.

Generated Caption (Culturally Irrelevant):
صورة تظهر منظرًا معماريًا قديمًا لمدينة ساحلية، تتضمن مسجدًا كبيرًا بقبته ومئذنته، محاطاً بأشجار النخيل ومباني منخفضة، مع البحر في الخلفية.
Translation: A coastal city view with a mosque, palm trees, and low-rise buildings by the sea.

Figure 1: Comparison of captions for the same image. The manual caption is culturally relevant, while the generated caption lacks cultural specificity.

1 Introduction

Image captioning, the automatic generation of natural language descriptions for visual content (Hossain et al., 2019), represents a fundamental challenge at the intersection of computer vision and natural language processing (Saraswat et al., 2024). While significant progress has been achieved for high-resource languages, particularly English, Arabic image captioning remains severely underexplored despite Arabic being spoken by over 400 million people worldwide (Mohamed et al., 2023b).

The challenges of Arabic image captioning extend beyond typical technical hurdles. Arabic's rich morphology, diverse dialectal variations, short vowel omissions, right-to-left script, and cultural diversity require specialized approaches that consider linguistic, cultural, and contextual factors (Jarrar et al., 2023b). Moreover, the lack of large-scale, high-quality Arabic image-caption datasets has hindered progress in this domain.

To address these challenges and highlight the unique issues in Arabic image captioning, we organized the ImageEval 2025 shared task, which comprised two complementary subtasks: Subtask

1, a collaborative image captioning datathon, and Subtask 2, an evaluation of Arabic image captioning models. The task design follows principles of cultural and linguistic authenticity, methodological diversity, and rigorous evaluation. Subtask 1 ensured that captions accurately reflected the perspectives and contextual norms of Arabic speakers, moving beyond direct translations from other languages. Figure 1 illustrates a comparison between manual and generated captions for the same image, where the manual caption reflects cultural context and historically specific information, whereas the generated caption (by GPT-5 mini) provides a general description with limited cultural relevance.

Subtask 2 encouraged participating teams to experiment with a broad range of modeling strategies, including zero-shot, few-shot, and fully supervised approaches. Model outputs were evaluated using a combination of widely adopted automatic metrics for image captioning, such as BLEU (Papineni et al., 2002) and cosine similarity (Sharif et al., 2020), as well as LLM-based assessments that capture semantic correctness and contextual appropriateness (Zhang et al., 2025). In addition, human evaluation was conducted to provide a complementary benchmark, focusing on fluency, cultural adequacy, and alignment with the visual content, thereby assessing subjective quality aspects not captured by automatic metrics. The shared task explicitly addresses challenges such as dataset scarcity, morphological complexity, cultural specificity, metric suitability, and resource constraints in Arabic NLP research.

This paper presents a comprehensive overview of ImageEval 2025, including our motivation, task design principles, data collection methodology, evaluation framework, baseline models, and analysis of participant approaches and results. Our contributions include:

- Introduce the first large-scale shared task for Arabic image captioning, combining collaborative data creation with competitive model evaluation.

- Comprehensive evaluation framework incorporating automatic metrics, LLM-based assessment, and human evaluation.

- Analysis of cultural and linguistic challenges specific to Arabic image captioning.

- Release all resources, including datasets, evaluation tools, and baseline models, as open-source.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 presents the shared task overview, Section 4 describes the evaluation methodology, Sections 5 and 6 detail Subtasks 1 and 2, Section 7 discusses challenges and insights, Section 8 covers impact and future directions, and Section 9 concludes the paper.

2 Related Work

2.1 Evolution of Image Captioning

Early image captioning relied on template-based (Farhadi et al., 2010; Kulkarni et al., 2013) and retrieval methods (Devlin et al., 2015; Ordonez et al., 2011), but the field was revolutionized by the adoption of encoder-decoder frameworks, where CNNs extract image features and RNNs or LSTMs generate captions (Stefanini et al., 2023; Ming et al., 2022; Hossain et al., 2018; Verma et al., 2023). The introduction of attention mechanisms allowed models to focus on salient image regions, improving caption relevance and fluency (Yu et al., 2019; Liu et al., 2020; Yan et al., 2021; Wang et al., 2020; Gao et al., 2020). Transformer-based models further advanced the field by enabling parallel processing and capturing long-range dependencies, leading to state-of-the-art results on benchmarks like MSCOCO (Yu et al., 2019; Yan et al., 2021; Xian et al., 2022; Parvin et al., 2023). Recent vision-language models such as CLIP, BLIP, and GPT-4V leverage large-scale pretraining and multimodal fusion, achieving remarkable performance and enabling new applications in accessibility and content retrieval (Khodave and Powar, 2025; Cho and Oh, 2023; Betala and Chokshi, 2024; Nguyen et al., 2023).

2.2 Multilingual and Cross-lingual Image Captioning

Multilingual image captioning has gained traction, with datasets like COCO-CN and Crossmodal-3600 supporting multiple languages (Cho and Oh, 2023; Li et al., 2019; Song et al., 2023). Most research, however, still focuses on resource-rich languages, with English dominating available data and benchmarks (Khodave and Powar, 2025; Cho and Oh, 2023; Li et al., 2019; Song et al., 2023). Cross-lingual transfer approaches, such as using visual pivots or synthetic data, have shown promise in generating captions for low-resource languages

(Al-Buraihy and Wang, 2024; Zhang et al., 2023; Hitschler et al., 2016; Song et al., 2023). Recent models employ transformer architectures and reinforcement learning to improve semantic and stylistic alignment across languages (Al-Buraihy and Wang, 2024; Zhang et al., 2023; Song et al., 2023). Despite these advances, morphologically complex languages like Arabic remain underexplored, and open-source models often lag behind proprietary systems in multilingual performance (Cho and Oh, 2023; Zha et al., 2022; Betala and Chokshi, 2024; Song et al., 2023).

2.3 Arabic NLP and Multimodal Processing

Arabic NLP presents unique challenges due to its rich morphology, complex script, and wide dialectal variation (Nayouf et al., 2023). Moreover, the meaning of many Arabic words can shift significantly depending on context (Jarrar, 2021; Akra et al., 2025). Several studies have been conducted on Arabic image captioning (Elbedwehy and Medhat, 2023; Emami et al., 2022b; ElJundi et al., 2020; Afyouni et al., 2021; Alsayed et al., 2023; Hejazi and Shaalan, 2021). While significant progress has been made in text-only Arabic NLP, multimodal applications, especially image captioning, are still nascent. Recent studies have proposed transformer-based and hybrid models for Arabic image captioning, often leveraging pre-trained language models such as AraBERT, MARBERT, and CamelBERT (Badarneh et al., 2025; Yu et al., 2019; Elbedwehy and Medhat, 2023; Emami et al., 2022b; Afyouni et al., 2021; Alsayed et al., 2023; Sabri, 2021). These models have demonstrated improved performance over translation-based approaches, but the lack of large, high-quality Arabic datasets remains a major bottleneck (Elbedwehy and Medhat, 2023; Emami et al., 2022b; ElJundi et al., 2020; Afyouni et al., 2021; Alsayed et al., 2023; Hejazi and Shaalan, 2021). Comparative studies highlight the importance of tailored preprocessing and feature extraction for Arabic, with some models achieving BLEU-4 scores up to 0.16, outperforming earlier work (Elbedwehy and Medhat, 2023; Alsayed et al., 2023; Sabri, 2021; Hejazi and Shaalan, 2021).

2.4 Shared Tasks in Multimodal NLP

Shared tasks and benchmarks such as MSCOCO, VQA, and COCO-CN have been instrumental in advancing image captioning and multimodal NLP (Khodave and Powar, 2025; Cho and Oh, 2023; Li et al., 2019; Betala and Chokshi, 2024). These challenges foster innovation, provide standardized evaluation, and drive the development of robust models (Khodave and Powar, 2025; Cho and Oh, 2023; Li et al., 2019; Betala and Chokshi, 2024). In addition, several Arabic shared tasks have addressed a range of NLP tasks, including named entity recognition (Jarrar et al., 2024, 2023a), language understanding (Khalilia et al., 2024), and dialect identification (Abdul-Mageed et al., 2024, 2023), demonstrating the value of community-driven evaluation across diverse language technologies. However, no major shared task has specifically targeted Arabic image captioning, highlighting a significant gap and an opportunity for future community-driven efforts (Cho and Oh, 2023; Betala and Chokshi, 2024; Sabri, 2021).

3 Shared-task Overview

The ImageEval 2025 shared task comprises two primary subtasks: Subtask 1, the Image Captioning Datathon, and Subtask 2, the Image Captioning Models Evaluation. Subtask 1 focuses on the manual creation of Arabic image captions, requiring participants to produce natural, culturally appropriate, and contextually aligned descriptions. Captions must be written manually, without the use of generative AI tools, and participants were provided with minimal contextual information about the images. This guidance helps teams generate meaningful captions that accurately reflect the content and cultural context of each image.

Subtask 2 evaluates the performance of Arabic image captioning models. Participants are allowed to use external datasets and retrieval-augmented generation (RAG) approaches; however, the submitted system must rely entirely on the provided dataset for evaluation. This requirement ensures a standardized and fair comparison across participating models.

The shared task received 44 registrations, and during the test phase, 8 teams submitted a total of 111 entries (109 submissions for Subtask 1 and 2 submissions for Subtask 1). In addition, 8 system description papers were submitted and all were accepted. To facilitate consistent evaluation and scoring of submissions, we employed Codabench[12], a well-established platform for shared-task evaluation. Furthermore, we established and shared a dedicated web page for the shared task, providing

[1] https://www.codabench.org/competitions/9447/
[2] https://www.codabench.org/competitions/9450/

participants with guidelines and detailed information as a reference [3]. Table 1 presents a detailed overview of the participating teams, listed in alphabetical order, along with their affiliations and the subtasks in which they participated.

4 Evaluation

4.1 Human Evaluation

We selected approximately 5% of the test data and applied four qualitative metrics to all participating teams. Each metric was rated on a scale from 1 (lowest) to 4 (highest):

- **Cultural Relevance** – Measures whether the description reflects cultural specificity and provides contextual information related to the scene.

- **Conciseness** – Assesses whether the description conveys information directly and succinctly, without unnecessary repetition or dispersion of details.

- **Completeness** – Evaluates the extent to which the description covers all aspects of the image, including events, entities, and relevant elements.

- **Accuracy** – Measures whether the description contains correct information, free from factual or conceptual errors.

4.2 Automatic Metrics

The task considered the following metrics for automatic evaluation of submissions.

- **BLEU** measures n-gram ($n \in [1, 4]$) overlap between generated and reference captions, and applies smoothing for sparse higher-order n-grams.

- **ROUGE scores**: (ROUGE-1, ROUGE-2, and ROUGE-L) are recall-oriented; they measure how many reference n-grams are recovered by the candidate caption and the longest common subsequence.

- **Cosine similarity**: compares the angular distance between vector representations of the captions. For this task, we used term-frequency–inverse-document-frequency (TF–IDF) vectors, where terms are n-grams ($n \in [1, 4]$) and each caption is a document.

- **Jaccard Similarity**: calculates the intersection over union of unique word sets, providing a set-based overlap measure.

- **Lin Similarity**: is an information-theoretic metric that computes twice the ratio of the information content (IC) of the least common subsumer of both captions, divided by the sum of the IC of both captions.

4.3 LLM as a Judge

We incorporated LLM as a judge in the scoring pipeline. Specifically, we employed the OpenAI GPT-4o model through its API, with a fixed random seed of 42, and an inference temperature of 0.0 to ensure reproducibility, using a task-specific system prompt (Appendix B).

For each (candidate, reference) caption pair, we provided a structured prompt and instructed the LLM to assign an integer score between 1 and 10, where 1 is lowest and 10 is highest similarity. The evaluation criteria emphasized semantic accuracy, relevance, and fluency of the candidate caption compared to the reference one.

Model outputs were parsed to reduce ambiguity, and evaluations were executed concurrently for efficiency. Final submission scores were obtained by averaging across all pairs and mapping results to a normalized $[0, 100]$ scale.

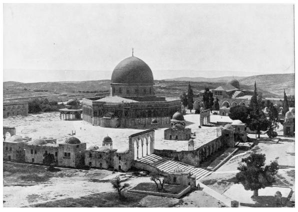
5 Subtask 1: Image Captioning Datathon

Images depict diverse visual scenes that require contextually rich and culturally informed descriptions, which motivated the Image Captioning Datathon (Subtask 1). This subtask aims to generate captions that are both linguistically natural and culturally appropriate for Arabic. Given an image I, the goal is to produce a caption C that accurately describes the content of I while reflecting Arabic language norms and cultural context. Participants were provided with a set of images and tasked with manually creating descriptive captions that emphasize meaning, context-awareness, and cultural grounding. Submissions were required in a CSV format, containing the corresponding image ID and the generated caption for each image in the test set. Figure 2 illustrates an example image along with its manually annotated caption.

[3] https://sina.birzeit.edu/image_eval2025/

Team	Affiliation	Subtask 1	Subtask 2
AZLU (Yassine et al., 2025)	Lebanese Univ., Birzeit Univ., Al Azhar Univ.	✓	
BZU-AUM (Alkhanafseh et al., 2025)	Birzeit Univ.	✓	
Averroes (Saeed et al., 2025)	Applied Innovation Center, Georgia Tech		✓
Phantom Troupe (Abu Horaira et al., 2025)	Chittagong Univ. of Engineering and Technology		✓
VLCAP (Elchafei and Fashwan, 2025)	Ulm Univ., Alexandria Univ.		✓
Codezone Research Group (Bichi et al., 2025)	Baba Ahmed Univ. Kano		✓
ImpactAi (Al-Qasem and Hendi, 2025)	Ggateway		✓
NU_Internship (Gaber et al., 2025)	Nile Univ., Ain Shams Univ., Alex. Univ.		✓

Table 1: Participating teams in ImageEval 2025 and their subtasks.



Arabic: صورة لساحة مسجد قبة الصخرة في الحرم الشريف

English Translation: An image of the courtyard of the Dome of the Rock Mosque in Al-Haram Al-Sharif.

Figure 2: Example image with corresponding caption.

5.1 Dataset

The dataset comprises $4,000$ open-source images collected from multiple domains with careful consideration to ensure cultural relevance and to avoid sensitive or inappropriate content. The dataset was systematically partitioned into 16 batches, each containing 250 images.

The images represent a broad spectrum of Palestinian cultural and social contexts. They encompass everyday life, the activities of liberation movements including military training, the lived experiences of refugees, and significant historical and touristic landmarks. The selection process prioritized diversity of perspectives to produce a dataset that is both rich and representative.

For evaluation, two batches (500 images) with pre-existing manual annotations were specified as mandatory. These annotations served as the reference ground truth for assessing the quality of the captions generated by participating teams. All teams were required to submit captions for these batches. In addition, teams were allowed to select further batches for annotation, provided that any chosen batch was captioned in its entirety.

5.2 Annotation Guidelines

Annotation guidelines were developed to ensure consistency across participants. Alongside these guidelines, an annotation file was provided containing 250 images organized into five sheets of 50 images each. Each sheet included a short contextual description, a thumbnail preview, and a URL to the original high-resolution image.

Participants were instructed to write captions in Modern Standard Arabic (MSA), avoiding colloquial or dialectal forms. Each caption was required to be between 15 and 100 words (ideally around 50 words, written in $3-4$ sentences). Captions were expected to be narrative in style, reflecting emotions, events, historical context, and cultural significance, rather than simply listing visible objects. To ensure quality and consistency, participants were required to perform all annotations manually without AI assistance and to develop their own detailed captioning guidelines for internal use.

5.3 Evaluation and Result

For Subtask 1, captions were evaluated using human evaluation (4.1), automatic metrics (4.2), and LLM as a judge (4.3). Since cosine similarity and LLM-based scores showed higher alignment with human evaluation, they were used for final ranking. The combined results are summarized in Table 2.

According to automatic metrics, BZU-AUM (Alkhanafseh et al., 2025) achieved the highest cosine similarity (65.53), while AZLU (Yassine et al., 2025) obtained the highest LLM Judge Score (41.53). Human evaluation results indicate BZU-AUM scored highest in cultural relevance (3.24) and completeness (3.08), whereas AZLU scored highest in conciseness (3.44) and accuracy (3.16).

The results indicate different annotation tendencies between the two teams, with BZU-AUM producing more complete and culturally relevant descriptions, while AZLU provided captions that were comparatively more concise and accurate.

5.4 Discussion

The teams approached manual captioning through varied annotation strategies, team composition, quality control practices, and cultural adaptation methods.

Teams	Automatic Evaluation				Human Evaluation			
	Rank (Cosine)	Cosine Similarity	Rank (LLM)	LLM Judge Score	Cultural Relevance	Conciseness	Completeness	Accuracy
AZLU	2	59.15	1	**41.53**	3.20	**3.44**	2.88	**3.16**
BZU-AUM	1	**65.53**	2	32.42	**3.24**	2.76	**3.08**	2.92

Table 2: Subtask 1 Results: Automatic Evaluation (Cosine Similarity, LLM Judge Score) and Human Evaluation (Cultural Relevance, Conciseness, Completeness, Accuracy).

Annotation strategies varied with emphasis on narrative richness and contextual detail, while the other team focused on brevity and precision. These tendencies are reflected in the completeness and cultural depth favoring one team versus prioritized conciseness and accuracy.

Team composition played a role in shaping annotation styles, as teams included native Arabic speakers with dialectal backgrounds and subject matter experts for historically or culturally sensitive images. Quality assurance reviewers were also engaged to enhance consistency.

Quality control measures centered on internal review processes to ensure that captions adhered to guidelines and maintained fluency. While *inter-annotator agreement was not systematically enforced across all teams*, they adopted informal checks for coherence and style alignment.

Cultural adaptation approaches were particularly important, as annotators sought to embed historical references, social practices, and cultural nuances in the captions. This emphasis helped maintain cultural relevance while ensuring captions extended beyond object description into meaningful narrative.

6 Subtask 2: Image Captioning Model Evaluation

Subtask 2 addresses the development of models for automatic Arabic image captioning. Given an image $i \in I$, the goal is to generate an Arabic caption c_i that is both *contextually accurate* and *culturally relevant*. Participants were provided with a curated dataset of manually annotated Arabic images, divided into training and test subsets. The training subset was shared for model development, while the test set was released later for caption generation. Submissions consisted of automatically generated captions C for each test image $i \in I$, and were evaluated against the ground truth captions using established automatic metrics (see Section 4.2) through Codabench.

6.1 Dataset

We prepared a curated dataset of $3,471$ manually annotated Arabic image-caption pairs, comprising $2,718$ images for training with ground-truth captions and 753 images reserved for final evaluation. The images capture a wide range of Palestinian cultural and social contexts, including everyday life, the activities of liberation movements such as military training, the experiences of refugees, and notable historical and touristic landmarks. The dataset is publicly available through Hugging Face[4][5].

6.2 Baselines

To establish performance benchmarks, we established two baselines on our human-annotated dataset: zero-shot and fine-tuning. The code for these baselines is publicly available on GitHub[6].

Zero-Shot Baseline

For the zero-shot baseline, we employed `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025), a vision–language model with a unified image encoder and autoregressive text decoder. The model was applied directly in inference mode, without any task-specific fine-tuning, to assess its ability to generate Arabic captions "out-of-the-box."

A multimodal prompt was designed to combine (i) the raw image and (ii) an instruction in Arabic guiding the model to generate culturally appropriate captions (15–50 words). The exact prompt template is provided in Appendix A.1. Inference was performed with a maximum generation length of 128 tokens. Outputs were collected in a structured format to facilitate evaluation. The final results are summarized in Table 3

Fine-Tuned Baseline

The same model was fine-tuned on the dataset using supervised fine-tuning (SFT) with LLAMA-FACTORY. Each training instance was formatted

[4] Train: `SinaLab/ImageEval2025Task2TrainDataset`

[5] Test: `SinaLab/ImageEval2025Task2TestDataset`

[6] Baselines: GitHub Repository

as a two-turn conversation in which the human prompt contained the image and a request for description in Arabic, and the assistant response was the corresponding gold caption.

Fine-tuning was carried out with parameter-efficient adaptation using LoRA. We trained for 15 epochs with a batch size of 16 and a maximum sequence length of 1024 tokens. Optimization employed AdamW with a learning rate of 2×10^{-5} and a cosine decay schedule with warmup. Dropout was set to 0.1, and the LoRA configuration used a rank of 8 with scaling parameter $\alpha = 16$. Training and validation losses were monitored throughout, and checkpoints were saved regularly. The fine-tuned model is publicly available[7].

Fine-tuning consistently improved performance across all evaluation metrics compared to the zero-shot baseline, as shown in Table 3, confirming the effectiveness of task-specific adaptation for Arabic image captioning.

Baseline	BLEU-1	BLEU-4	Cosine Similarity	LLM Judge (%)
Zero-shot	0.0992	0.0133	0.5577	27.11
Fine-tuned	0.1698	0.0305	0.5846	30.82

Table 3: Baseline performance on the dataset.

6.3 Participant Systems

All teams adapted pretrained vision–language models. They relied on translation, fine-tuning, or augmentation, and differed in how they restructured the captioning pipeline.

Averroes (Saeed et al., 2025) employs a two-stage pipeline, where one Qwen2.5-VL-7B based model generates detailed descriptions and another refines captions. They augmented training data with AyaVision8B and used BLEU scores to validate and pair with randomized image transformations. Its key contribution is systematic augmentation that enhances diversity without distorting the data distribution.

Codezone Research Group (Bichi et al., 2025) uses a zero-shot translation pipeline: BLIP generates English captions, which are translated to Arabic with M2M100. To ensure consistency in evaluation, the output is normalized by removing diacritics, Tatweel, and punctuation. Unlike others, it avoids fine-tuning, showcasing the viability of off-the-shelf models combined with robust translation.

ImpactAi (Al-Qasem and Hendi, 2025) proposed a region-aware captioning method based on the Region Features Transformer (CRAFT). The approach extracts a set of salient regions from each image using Faster R-CNN and encodes these region features through a transformer encoder–decoder architecture, paired with ArabGloss-BERT tokenization. This integration distinguishes it from other submitted methods.

Phantom Troupe (Abu Horaira et al., 2025) uses a translation-centered pipeline: Arabic captions are translated to English with the Qwen3-14B model for training and then back-translated at inference. They fine-tune Qwen2.5-VL-7B with LoRA for efficient adaptation. Its distinctive feature is the preservation of cultural nuances during translation while leveraging strong English captioning models.

NU_Internship (Gaber et al., 2025) adapted a vector store-based approach to enhance domain adaptability. They used Gemini-2.5 Flash, expanded the training data, and experimented with both zero-shot and fine-tuning, with and without RAG. To fuse the outputs of the top-performing models, they applied a meta-learning stacked ensemble using an LLM, selection guided by BLEU and cosine similarity metrics.

VLCAP (Elchafei and Fashwan, 2025) VLCAP is an Arabic image captioning framework that conditions generation on interpretable visual labels. A hybrid vocabulary is derived by extracting noun-like keywords from training dataset captions and augmenting them with over 21K translated Visual Genome concepts. Three retrieval experiments are conducted using mCLIP, AraCLIP, and Jina V4, where the top-k most relevant labels for each image are identified to construct the Arabic prompt for captions generation. These prompts, together with the original image, are provided to Qwen-VL and Gemini Pro Vision in separate settings. The best results are achieved when combining mCLIP for label retrieval with Gemini Pro Vision for caption generation, producing culturally coherent and contextually accurate Arabic captions, while AraCLIP with Qwen-VL excels in human-judged quality.

6.4 Evaluation and Results

Subtask 2 was assessed using the same three perspectives introduced earlier: automatic metrics (4.2), LLM as a judge (4.3), and human evaluation (4.1). Table 4 presents the comparative results for all participating teams.

VLCAP scored the highest cosine similarity

[7] Finetuned Baseline: Hugging Face

Teams	Automatic Evaluation				Human Evaluation			
	Rank (Cosine)	Cosine Similarity	Rank (LLM)	LLM Judge Score	Cultural Relevance	Conciseness	Completeness	Accuracy
Averroes	2	58.55	1	**33.97**	**3.63**	**3.43**	2.60	2.80
Codezone Research Group	6	38.30	6	15.14	1.10	2.03	1.47	2.03
ImpactAi	4	56.22	4	26.55	3.13	2.73	1.77	1.97
NU_Internship	5	55.32	5	24.87	2.57	2.97	2.13	2.23
Phantom Troupe	3	57.48	3	31.43	3.40	3.27	2.33	2.40
VLCAP	1	**60.01**	2	33.05	2.57	3.17	**2.67**	**2.97**

Table 4: Subtask 2 Results: Automatic Evaluation (Cosine Similarity, LLM Judge Score) and Human Evaluation (Cultural Relevance, Conciseness, Completeness, Accuracy).

(60.01), whereas Averroes ranked top with the LLM-Judge (33.97). Human evaluation highlights further distinctions: Averroes led in cultural relevance (3.63) and conciseness (3.43), while VLCAP ranked highest in completeness (2.67) and accuracy (2.97). Phantom Troupe also performed strongly, particularly in cultural relevance and conciseness.

6.5 Discussion

Submissions varied across model architectures, training, and fine-tuning strategies.

Model architectures were largely based on fine-tuning pretrained multilingual vision–language models, often with LoRA adapters for efficiency. Several teams relied on cross-lingual transfer by generating English captions and translating to Arabic, while one system introduced region-aware modeling with custom transformer components.

Training strategies included data augmentation, where image transformations and caption validation expanded the training set, and multi-stage pipelines that first produced detailed image descriptions before refining them into captions.

Arabic-specific optimizations focused on cultural and linguistic nuances during translation, dedicated Arabic tokenizers, and normalization to improve consistency in evaluation.

7 Challenges and Insights

Arabic image captioning faces significant challenges that stem from linguistic, cultural, and resource-related gaps. Unlike English, where large-scale datasets and robust models exist, Arabic research suffers from a severe shortage of high-quality, publicly available datasets (Emami et al., 2022a; Attai and Elnagar, 2020; Mohamed et al., 2023a; Kadaoui et al., 2025). Most available resources are translations from English rather than native Arabic captions, which fail to capture authentic linguistic patterns and cultural nuances (Ibrahim et al., 2025). This scarcity not only limits standardized benchmarking but also fragments research efforts, as scholars are forced to build small-scale datasets in isolation. Beyond resource limitations, Arabic itself introduces unique challenges due to its morphological richness, right-to-left script, connected character system (where OCR is needed), and extensive dialectal variation. These features make direct transfer of English-based methods ineffective, while translation-based approaches accumulate errors and degrade caption quality (Attai and Elnagar, 2020; Mohamed et al., 2023a). Cultural representation further complicates the task, as most image datasets are Western-centric and fail to reflect Arab cultural contexts, leading to mismatches between images and captions (Attai and Elnagar, 2020; Al-Buraihy et al., 2025). Addressing these challenges requires not only technical advances in preprocessing and modeling but also the creation of culturally authentic datasets tailored to Arabic's linguistic and social complexity.

ImageEval 2025 contributes to the benchmarking and further development of Arabic image captioning, offering a common ground for system comparison and incremental progress. By releasing the datasets, baseline models, and evaluation tools, this shared task aims to support the community and facilitate future research in Arabic multimodal NLP.

8 Future Directions

Building on the success and insights from ImageEval 2025, we identify several promising directions for future research and development in Arabic image captioning.

Future research should prioritize the development of evaluation metrics that more effectively capture Arabic morphological complexity, cultural nuances, and semantic variability, addressing limitations of current automatic measures. Addressing

Arabic dialectal diversity is another critical area, requiring models capable of adapting to regional linguistic variations and code-switching phenomena. Furthermore, enhancing cross-lingual transfer learning from high-resource languages while maintaining Arabic linguistic and cultural fidelity represents an important methodological challenge.

Efforts should be directed toward scaling data collection to develop larger and more diverse Arabic multimodal datasets that include additional domains and cultural contexts. Furthermore, translating these research advances into practical applications can enhance accessibility, content management, and educational technologies for Arabic-speaking communities.

The foundations established through ImageEval 2025 provide a robust platform for these future endeavors, with open-source resources and established methodologies enabling continued progress in Arabic multimodal NLP research.

9 Conclusion

ImageEval 2025 represents a significant milestone in Arabic multimodal NLP, addressing the critical gap in Arabic image captioning through innovative task design and community collaboration. The shared task successfully created valuable resources for the research community while highlighting unique challenges and opportunities in Arabic multimodal processing.

Our dual-task approach combining collaborative data creation with competitive model evaluation proved effective in both advancing the state-of-the-art and fostering community engagement. The results demonstrate both the challenges inherent in Arabic image captioning and the potential for significant progress through focused research efforts.

The datasets, evaluation tools, and insights generated through ImageEval 2025 provide a foundation for continued research in Arabic multimodal NLP. We anticipate that this work will catalyze further developments in Arabic vision-language processing and contribute to more inclusive and culturally aware AI systems.

Limitation

This study is limited to Palestinian cultural representation and does not cover other Arabic-speaking regions. The dataset captions are exclusively in MSA and do not include regional dialects. Therefore, while suitable for training models on Palestinian cultural contexts, the dataset's applicability to other Arabic cultures is restricted. Expanding to additional dialects and regions is necessary to enable broader cultural generalization in model training.

Acknowledgments

We thank all participating teams for their valuable contributions to Arabic multimodal NLP. We also acknowledge the organizers of the Arabic NLP 2025 Conference and the broader Arabic NLP community for their support. Special thanks to our annotation quality reviewers and cultural consultants who ensured the authenticity and appropriateness of our guidelines and datasets. This shared task was partially funded by the NPRP14C-0916-210015 grant, awarded by the Qatar Research, Development and Innovation Council (QRDI), and was also partially funded by the Research Committee at Birzeit University (2025T2376).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. Nadi 2023: The fourth nuanced arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Abu Horaira, Farhan Amin, Sakibul Hasan, Md. Tanvir Ahammed Shawon, and Muhammad Ibrahim Khan. 2025. Phantomtroupe at imageeval shared task: Multimodal arabic image captioning through translation-based fine-tuning of llm models. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Imad Afyouni, Imtinan Azhar, and Ashraf Elnagar. 2021. Aracap: A hybrid deep learning architecture for arabic image captioning. In *Procedia Computer Science*, pages 382–389.

Diyam Akra, Tymaa Hammouda, and Mustafa Jarrar. 2025. Quranmorph: Morphologically annotated quranic corpus. Technical report, Birzeit University.

Emran Al-Buraihy and Dan Wang. 2024. Enhancing cross-lingual image description: A multimodal approach for semantic relevance and stylistic alignment. *Computers, Materials & Continua*.

Emran Al-Buraihy, Dan Wang, Tariq Hussain, R. Attar, A. Alzubi, Khalid Zaman, and Zengkang Gan. 2025. Aratraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging. *Scientific Reports*.

Rabee Al-Qasem and Mohannad Hendi. 2025. Impactai at imageeval 2025 shared task: Region-aware transformers for arabic image captioning—a case study on the palestinian narrative. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Mohammed Alkhanafseh, Ola Surakhi, and Abdallah Abedaljalill. 2025. Bzu-aum at imageeval 2025: An arabic image captioning dataset for conflict narratives with human annotation. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Ashwaq Alsayed, Thamir M. Qadah, and Muhammad Arif. 2023. A performance analysis of transformer-based deep learning models for arabic image captioning. *Journal of King Saud University – Computer and Information Sciences*, 35(8):101684.

Anfal Attai and Ashraf Elnagar. 2020. A survey on arabic image captioning systems using deep learning models. *International Conference on Innovations in Information Technology*.

Israa Al Badarneh, Rana Husni Al Mahmoud, Bassam H. Hammo, and Omar S. Al-Kadi. 2025. Attention-based transformer model for arabic image captioning. *Neural Computing and Applications*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Siddharth Betala and Ishan Chokshi. 2024. Brotherhood at WMT 2024: Leveraging LLM-generated contextual conversations for cross-lingual image captioning. *arXiv preprint arXiv:2409.15052*.

Abdulkadir Shehu Bichi, Ismail Dauda Abubakar, Fatima Muhammad Adam, Aminu Musa, Auwal Umar Ahmed, Abubakar Ibrahim, Khadija Salihu Aua, Aisha Mustapha Ahmed, and Mahmud Said Ahmed. 2025. Codezone research group at imageeval 2025 shared task: Arabic image captioning using BLIP and M2M100—a two-stage translation approach. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Suhyun Cho and Hayoung Oh. 2023. Generalized image captioning for multilingual support. *Applied Sciences*.

Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.

Samar Elbedwehy and Tamer Mohammed Ibrahim Medhat. 2023. Improved arabic image captioning model using feature concatenation with pre-trained word embedding. *Neural Computing and Applications*, 35:19051–19067.

Passant Elchafei and Amany Fashwan. 2025. Vlcap at imageeval 2025 shared task: Multimodal arabic captioning with interpretable visual concept integration. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Obeida ElJundi, Mohamad Dhaybi, Kotaiba Mokadam, Hazem Hajj, and Daniel Asmar. 2020. Resources and end-to-end neural network models for arabic image captioning. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pages 233–241.

Jonathan Emami, P. Nugues, A. Elnagar, and Imad Afyouni. 2022a. Arabic image captioning using pre-training of deep bidirectional transformers. *International Conference on Natural Language Generation*.

Jonathan Emami, Pierre Nugues, Ashraf Elnagar, and Imad Afyouni. 2022b. Arabic image captioning using pre-training of deep bidirectional transformers. In *Proceedings of the 15th International Conference on Natural Language Generation*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer.

Rana Gaber, Seif Eldin Amgad, Ahmed Sherif Nasri, Mohamed Ibrahim Ragab, and Ensaf Hussein Mohamed. 2025. NU_Internship team at imageeval 2025: From zero-shot to ensembles—enhancing grounded arabic image captioning. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. 2020. Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:1112–1131.

Hani 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*.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Md. Zakir Hossain, Ferdous Sohel, Mohd. Fairuz Shiratuddin, and Hamid Laga. 2018. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51:1–36.

Md. Zakir Hossain, Ferdous Sohel, Mohd. Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

George Ibrahim, Rita Ramos, and Yova Kementchedjhieva. 2025. Concap: Seeing beyond english with concepts retrieval-augmented captioning. *arXiv.org*.

Mustafa Jarrar. 2021. The arabic ontology: An arabic wordnet with ontologically clean content. *Applied Ontology*, 16(1):1–26.

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa' Omar. 2023a. Wojoodner 2023: The first arabic named entity recognition shared task. In *Proceedings of ArabicNLP 2023*, pages 748–758, Singapore (Hybrid). Association for Computational Linguistics.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Wojoodner 2024: The second arabic named entity recognition shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 847–857, Bangkok, Thailand. Association for Computational Linguistics.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023b. Lisan: Yemeni, iraqi, libyan, and sudanese arabic dialect corpora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7, Egypt. IEEE.

Karima Kadaoui, Hanin Atwany, Hamdan Hamid Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2025. Jeem: Vision-language understanding in four arabic dialects. *arXiv.org*.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. Arabicnlu 2024: The first arabic natural language understanding shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 361–371, Bangkok, Thailand. Association for Computational Linguistics.

Nikhil Gopal Khodave and Prathamesh S. Powar. 2025. Survey on multimodal image captioning approaches: Addressing contextual understanding, cross-dataset generalization, and multilingual captioning. *International Journal of Advanced Research in Science, Communication and Technology*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21:2347–2360.

Maofu Liu, Lingjun Li, Huijun Hu, Weili Guan, and Jing Tian. 2020. Image caption generation with dual attention mechanism. *Information Processing & Management*, 57:102178.

Yue Ming, Nannan Hu, Chunxiao Fan, Fan Feng, Jiangwan Zhou, and Hui Yu. 2022. Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA Journal of Automatica Sinica*, 9:1339–1365.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Alcoba Inciarte, and M. Abdul-Mageed. 2023a. Violet: A vision-language model for arabic image captioning with gemini decoder. *ARABICNLP*.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. 2023b. Violet: A vision-language model for arabic image captioning with gemini decoder. *arXiv preprint arXiv:2311.08844*.

Amal Nayouf, Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian arabic dialects with morphological annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of EMNLP 2023*, pages 12–23, Singapore. Association for Computational Linguistics.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems*, 24.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Hashem Parvin, Ahmad Reza Naghsh-Nilchi, and Hossein Mahvash Mohammadi. 2023. Transformer-based local-global guidance for image captioning. *Expert Systems with Applications*, 223:119774.

Sabri Monaf Sabri. 2021. Arabic image captioning using deep learning with attention. Master's thesis, University of Georgia.

Mariam Saeed, Sarah Elshabrawy, Abdelrahman Hagrass, Mazen Yasser, and Ayman Khalafallah. 2025. Averroes at imageeval 2025 shared task: Advancing arabic image captioning with augmentation and two-stage generation. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Mala Saraswat, Challa Vivekananda Reddy, and Garandal Yashwanth Singh. 2024. Image captioning using NLP. In *2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP)*, pages 549–553. IEEE.

Naeha Sharif, Lyndon White, Mohammed Bennamoun, Wei Liu, and Syed Afaq Ali Shah. 2020. Wembsim: A simple yet effective metric for image captioning. *arXiv*.

Zijie Song, Zhenzhen Hu, Yuanen Zhou, Ye Zhao, Richang Hong, and Meng Wang. 2023. Embedded heterogeneous attention transformer for cross-lingual image captioning. *IEEE Transactions on Multimedia*, 26:9008–9020.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2023. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:539–559.

Akash Verma, Arun Kumar Yadav, Mohit Kumar, and Divakar Yadav. 2023. Automatic image caption generation using deep learning. *Multimedia Tools and Applications*, 83:5309–5325.

Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, Dagan Feng, and Tieniu Tan. 2020. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*, 98.

Tiantao Xian, Zhixin Li, Canlong Zhang, and Huifang Ma. 2022. Dual global enhanced transformer for image captioning. *Neural Networks*, 148:129–141.

Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, An-An Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. 2021. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:43–51.

Sarah Yassine, Sara Mahrous, and Rawan Sous. 2025. Azlu at imageeval 2025: Bridging linguistic and cultural gaps in arabic image captioning. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:4467–4480.

Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2022. Context-aware visual policy network for fine-grained image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:710–722.

Jing Zhang, Dan Guo, Xun Yang, Peipei Song, and Meng Wang. 2023. Visual-linguistic-stylistic triple reward for cross-lingual image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20:1–23.

Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. 2025. Crowd comparative reasoning: Unlocking comprehensive evaluations for LLM-as-a-judge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5059–5074, Vienna, Austria. Association for Computational Linguistics.

A Methodology

A.1 Zero-Shot Approach

The zero-shot methodology implements direct inference using the pre-trained QWEN2.5-VL-7B-Instruct model without domain-specific fine-tuning, serving as a crucial baseline for Arabic image captioning performance evaluation. The system loads the base model in `bfloat16` precision with automatic device mapping to optimize computational efficiency while maintaining model performance.

A.1.1 Prompt Engineering Strategy

A multimodal prompt was designed combining (i) the raw image and (ii) an instruction asking the model to generate a natural, culturally appropriate caption in Arabic (15–50 words) task within the Palestinian Nakba and Israeli occupation framework:

> *"You are an expert in visual scene understanding and multilingual caption generation. Analyze the content of this image, which is potentially related to the Palestinian Nakba and Israeli occupation of Palestine, and provide a concise and meaningful caption in Arabic - about 15 to 50 words. The caption should reflect the scene's content, emotional context, and should be natural and culturally appropriate. Do not include any English or metadata — The caption must be in Arabic."*

This design leverages the model's pre-trained knowledge about historical events, cultural sensitivity, and multilingual generation capabilities without requiring additional training data.

A.1.2 Inference Pipeline

The system utilizes the processor's chat template functionality for correct input formatting, followed by vision information processing for image data handling. Generation parameters are set with a maximum of 128 new tokens to ensure concise yet meaningful Arabic descriptions while preventing overly verbose outputs.

A.1.3 Methodological Advantages

This zero-shot approach provides several key advantages:

- **Rapid deployment** without training overhead
- **Unbiased evaluation** of pre-trained capabilities
- **Performance baseline** establishment for fine-tuned variant comparison
- **Domain-specific assessment** of cultural sensitivity and historical context understanding in Arabic

The systematic processing and structured CSV output enable comprehensive performance analysis across multiple evaluation metrics, supporting both quantitative assessment through BLEU scores and qualitative evaluation through LLM-as-a-judge scoring systems.

A.2 Fine-tuning Approach

A.2.1 Base Model

We fine-tune `Qwen/Qwen2.5-VL-7B-Instruct`, a vision–language model (VLM) with a unified image encoder and autoregressive text decoder. Supervised fine-tuning (SFT) is performed using LLAMA-FACTORY.

A.2.2 Task Formulation

The objective is Arabic image captioning. Each training example is a two-turn conversation:

- **Human**: "`<image>` Describe this image in Arabic."
- **Assistant**: gold Arabic description.

The dataset template `qwen2_vl` is used so that images and text are tokenized consistently with the base model.

A.2.3 Data Preparation

Source annotations are provided in an Excel file with `File Name` and `Description` columns. We convert rows to the LLaMA-Factory JSON format with absolute image paths:

- `conversations`: the prompt/response pair above.
- `images`: list with one absolute path to the corresponding JPEG.

Before training, we verify the existence and integrity of each image via `PIL.Image.verify()` and report any missing files.

A.2.4 Training Configuration

Parameter-efficient fine-tuning is applied with LoRA:

- **Stage**: SFT **Finetuning**: `lora` on all target modules.
- **LoRA hyperparameters**: rank r=8, α=16, dropout 0.1.
- **Sequence length**: cutoff = 1024 tokens.
- **Batching**: per-device batch size = 1, gradient accumulation = 16 (effective batch size 16).
- **Optimization**: AdamW with learning rate 2×10^{-5}, cosine LR schedule, warmup ratio 0.1.
- **Epochs**: 15.

A.2.5 Logging and Checkpointing

Training is launched via `llamafactory-cli train` with YAML configuration. We log every 5 steps and save checkpoints every 25 steps to the specified output directory. Loss curves are recorded for monitoring; external evaluators are not invoked in this pipeline.

A.2.6 Reproducibility Notes

We confirm tokenizer compatibility with Arabic text and report vocabulary size prior to training. All paths are absolute to avoid path-resolution errors during multi-process loading. The entire procedure is available at https://github.com/SinaLab/ImageCaptionSharedTask2025.

B LLM As a Judge System Prompt

```
You are an expert AI evaluator specializing in Arabic language and semantics. Your task is to act
as an impartial judge and evaluate the quality of a "model-generated caption" of a given image by
comparing it to a "ground truth caption" for the same image. You will not see the image itself.
Your entire evaluation must be based on the textual comparison of the two provided Arabic captions.
Assume the "ground truth caption" is the accurate and correct description of the image.
Evaluation Criteria: Please evaluate the "model-generated caption" based on the following criteria,
using a scale of 1 to 10, where 1 is Very Poor and 10 is Excellent.
Semantic Similarity: - How closely does the model's caption convey the same core meaning as the
ground truth? - Does the caption mention the same key objects, attributes, and actions as the ground
truth? Score 10: The meaning is identical or nearly identical. Score 1: The meaning is completely
different or irrelevant.
REPLY WITH THE SCORE ONLY. NO EXPLANATION
Caption to Evaluate:
```