AZLU at ImagEval Shared Task: Bridging Linguistics and Cultural Gaps in Arabic Image Captioning

Sarah Yassine[*]
Lebanese University / Lebanon
sarah.yassine.2@st.ul.edu.lb

Sara Mahrous
Al Azhar University / Egypt
mahroussara299@gmail.com

Rawan Sous
Birzeit University / Palestine
1200129@student.birzeit.edu

Abstract

Image captioning is the task of automatically generating natural language descriptions for visual content, with applications in search, social media, and beyond. While English captioning has advanced significantly, Arabic captioning remains underdeveloped due to a scarcity of high-quality, culturally relevant datasets. This work, conducted under the ImageEval 2025 Shared Task, addresses this gap by introducing a novel, manually annotated, open-source dataset for Arabic image captioning. Our curated resource consists of 500 unique black-and-white historical photographs documenting pivotal events in modern Palestinian and Lebanese history. The dataset spans from the British Colonial era in Palestine through the events of 1948, and includes documentation of the 1982 Israeli invasion of Beirut. This contribution provides a foundational resource to advance research in Arabic NLP and multimodal systems, offering a vital benchmark for models processing complex historical, cultural, and traumatic imagery.

1 Introduction

Despite significant progress in English image captioning, Arabic captioning remains an understudied challenge due to the language's complex morphology, dialectal diversity, and cultural nuances. These linguistic and contextual gaps hinder the development of robust captioning systems for Arabic content. A key issue identified in recent work on Arabic image captioning is the lack of a well-structured, high-quality dataset in Modern Standard Arabic (MSA) (Mohamed et al., 2023).

To address this gap, we participated in the ImageEval 2025 Shared Task (Bashiti et al., 2025), curating a high-quality dataset of 500 manually annotated images. Our approach enforces strict guidelines: captions are written exclusively in MSA, adhere to length constraints, and prioritize cultural relevance. This ensures consistency and broad usability for both native speakers and learners, while providing a reliable resource for fine-tuning Arabic-capable LLMs.

Crafting high-quality captions in Modern Standard Arabic (MSA) required meticulous precision to accurately describe culturally specific elements (e.g., Palestinian villages, Beirut streets). While regional dialects are prevalent, we employed MSA to ensure broad comprehensibility and establish a formal benchmark. Resources like The Living Arabic Project were leveraged to ensure linguistic accuracy and contextual relevance.

The system achieved strong performance, excelling in automated (41.53 LLM judge score) and human evaluations for conciseness (3.44) and accuracy (3.16). While semantic alignment was lower (59.15 cosine similarity), the methodology proved effective for generating concise, culturally and linguistically accurate Arabic captions.

In summary, our work makes the following contributions:

- A high-quality MSA caption dataset: Carefully annotated with no dialectal influence.
- Culturally relevant descriptions: Each caption reflects the cultural context of the image.
- Linguistically robust system: Our approach ensures factual and grammatical correctness

[*] Corresponding author.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 438–442
November 8-9, 2025 ©2025 Association for Computational Linguistics

in generated captions.

2 Background

2.1 Task Setup

For the ImageEval 2025 Shared Task (Subtask 1: Image Captioning Datathon), our methodology was designed to generate captions that explicitly address the primary evaluation criteria of linguistic quality and cultural relevance.

Input: Our input consisted of a collection of 500 uncaptioned images depicting Palestinian heritage (e.g., traditional villages, daily life) and Beirut during pivotal historical moments (e.g., the Lebanese Civil War, Israeli invasions).

Output: Our generated captions were designed to be culturally grounded in MSA, adhering to the following key constraints. First, a strict word count of between 10 and 50 words. Second, linguistic rigor was maintained by permitting only MSA, excluding regional dialect variants(e.g., اللوح/teaching board/ (al-lawh)), instead of (e.g., السبورة/blackboard/ as-sabbūrah, a regional colloquialism) (Za'ter and Talafha, 2022). Third, we mandated cultural precision. This required the use of specific, contextually appropriate vocabulary, as in the following example: صورة من داخل القدس تظهر المقوش والهندسة الداخلية للمدينة التي تعج بأهلها الذين يرتدون لباسهم التقليدي من عقال وحطة/ (*šūra min dāhil al-quds tazhar al-maqwash wa-l-handasa al-dāhiliyya li-l-madina allati ta'ij bi-ahlihā alladh*^{i}*na yartad*^{u}*na libāsahum al-taql*^{i}*d*^{i} *min 'iqāl wa-a*a) / (A view from Jerusalem's Old City showing its arched alleys and bustling crowds adorned in traditional keffiyehs and headbands).

2.2 Dataset Details

We manually annotated all 500 images with captions adhering to the above constraints. The dataset is divided into two primary batches of 250 images, each further split into four thematic subsets of 50 images.

2.3 Track Participation

Our participation in Subtask 1 (Image Captioning Datathon) (Bashiti et al., 2025) aimed to demonstrate high-quality caption generation under strict linguistic and cultural constraints. We contributed a rigorously annotated dataset to benchmark MSA compliance and highlight the importance of accurate historical and cultural context over generic descriptions.

3 System Overview

3.1 Design Rationale and Core Principles

Our captioning system addresses critical gaps in prior Arabic image captioning research. Previous work suffered from a scarcity of high-quality public datasets in MSA (Za'ter and Talafha, 2022), a tendency to generate generic descriptions lacking culturally significant details (Emami et al., 2022), and complications from Arabic's dialectal diversity that hinder linguistic consistency (Emami et al., 2022).

To overcome these issues, we implemented a controlled framework. Our primary objective was to create a high-quality, open-source MSA dataset to directly address its scarcity. Consequently, we enforced strict MSA usage, which involved replacing dialectal terms (e.g., سطل, *satl* 'bucket') with their MSA equivalents (دلو, dalw, Bucket).

3.2 Annotation Guidelines

To ensure consistency and quality, all annotators adhered to a strict set of rules.

First, the Language Standard required captions to be written exclusively in MSA, prohibiting dialectal terms to prevent linguistic interference.

Second, the Descriptive Depth guideline mandated comprehensive narratives of 10–50 words, avoiding simple labels (e.g., مسجد, masjid, mosque).

Third, Content Requirements obliged annotators to contextualize scenes by describing precise locations (e.g. المسجد الأقصى في البلدة القديمة بالقدس (*al-Masjid al-Aqsā fī al-balda al-qadīma bi-al-Quds*) 'Al-Aqsa Mosque in the Old City of Jerusalem'), actions, and cultural significance.

Fourth, standardized terminology was achieved by requiring annotators to source all terms from a project glossary validated using The Living Arabic Project dictionary (Living Arabic Project, n.d.).

Finally, Cultural and Historical Accuracy was ensured by verifying culturally significant terms (e.g. يافا (*Yāfā*) 'Jaffa') against historical and multilingual sources, including Wikipedia and digital archives.

3.3 Annotation Challenges and Resolution Strategies

The annotation process encountered several challenges, resolved through structured protocols: First, dialectal interference arises from the team's diverse dialects (e.g., using شروال, shirwāl for trousers). This was mitigated by developing a collaborative glossary to reach consensus on standard MSA terms (e.g. بنطلون, bantalōn, Pants). Second, politically and Culturally Sensitive Terminology required precise language for historical scenes. A mandatory consultation process with historical advisors was instituted to standardize terms (e.g. مجزرة, (majzara), massacre; النضال الفلسطيني (*al-nidāl al-filastīnī*) 'the Palestinian struggle'; الإحتلال البريطاني (*al-ihtilāl al-britānī*) 'the British occupation'). Third, Ambiguity in Transliterated Toponyms was addressed by implementing a verification protocol cross-referencing official maps and historical documents to confirm modern standard Arabic terms (e.g. طبريا (*Tabariyyā*) 'Tiberias').

3.4 Quality Assurance and Validation

A multistage validation process ensured the accuracy of captions. Geographic landmarks were verified against contemporary images from Wikipedia and official records. A linguist also performed random spot checks on finalized captions to verify adherence to all guidelines in Section 3.2.

4 Dataset

4.1 Dataset Overview

This dataset is based on resources from the Shared Task organizers, which we have significantly extended.

First, the organizers provided a core set of 500 uncaptioned images, each with a basic contextual note. The images were organized into ten thematic groups of 50 images.

Second, our contribution was to transform this into a vision and language dataset. We manually authored a relevant and descriptive caption in MSA for eachmage.

Finally, the complete data set contains 500 image caption pairs. Thematic coverage includes: Palestinian resistance (40%), Palestinian cities and villages (30%), events from the Palestinian-Zionist conflict (10%), Beirut during the Lebanese Civil War (10%), and Palestinian daily life and culture (10%).

4.2 Linguistic Analysis of Captions

To characterize the linguistic properties of our manually authored captions, we conducted a quantitative analysis of lexical and syntactic features. This provides a clear profile of the dataset for future users.

Lexical Diversity and Terminology: A frequency analysis of the corpus confirms its thematic focus. The most frequently named entities are location names, led by فلسطين (*Filastīn*) (Palestine, occurring in 32% of captions), القدس (*al-Quds*) (Jerusalem, 28%), and المسجد الأقصى (*al-Masjid al-Aqsā*) (Al-Aqsa Mosque, 15%). As the dataset documents historical events, conceptual nouns such as شهداء (*shuhadā'*) (martyrs) and مجازر (*majāzir*) (massacres) are also highly prevalent, appearing in approximately 35% and 25% of all captions, respectively. The name بيروت (*Bayrūt*) (Beirut) occurs in roughly 10% of the captions, aligning with its defined thematic share.

Syntactic Properties: The captions vary in length from 8 to 50 words, with an average length of 15 words, providing substantive descriptions. A manual analysis of a 100-caption sample revealed that approximately 60% utilize a nominal sentence structure (e.g., الجملة الاسمية (*al-jumla al-ismiyya*), which is typical of descriptive Arabic. Furthermore, given the historical nature of the images, the past tense is the predominant verbal form, used in over 80% of captions that contain a verb.

4.3 Quality Assurance Framework

To ensure the quality and reliability of the captions, we employed a multi-faceted evaluation strategy using automated metrics and human assessment. A detailed analysis of the evaluation results is presented in Section 5 (Results).

4.4 External Tools/Libraries

To ensure factual accuracy, toponym spellings and historical context were verified using Wikipedia and digital archives. This standardized Arabic transliteration against alternative names (e.g., يافا for Jaffa/Yafo), prevents ambiguous geographic references.

5 Results

5.1 Quantitative Results: Ranking and Performance of Official Metrics

In Subtask 1 of ImageEval 2025, performance was assessed through three complementary approaches:

First, Semantic Alignment (Cosine Similarity): Captions were evaluated by computing the average pairwise cosine similarity between TF-IDF vectorized character 3-grams of candidate and reference texts after text normalization. This metric quantifies lexical overlap, accounting for Arabic morphological variation. As shown in Table 1, BZU-AUM led (65.53), while our team (AZLU) scored 59.15.

Second, Automated Quality Assessment (LLM as Judge): A GPT-4o model evaluated captions on a 0–100 scale for semantic accuracy, fluency in MSA, and cultural relevance. Using fixed parameters and a structured prompt ensured reproducibility. Our team (AZLU) led this metric (41.53), reflecting strengths in coherence and relevance, while BZU-AUM scored 32.42.

Third, Manual Evaluation: A 5% stratified sample was evaluated by native Arabic speakers on four qualitative metrics (rated 1–4): Cultural relevance, Concise, Completeness, and Accuracy.

Table 1

Rank (Cosine)	Participants	Cosine Similarity Mean	Rank (LLM)	LLM Judge Score
1	BZU	65.53	2	32.42
2	AZLU	59.15	1	41.53

Table 1: Cosine Similarity and LLM Judge Score results for participating teams.

Table 1 reveals a performance inversion: BZU-AUM led in Cosine Similarity (65.53) but scored lower on the LLM Judge (32.42), while our team (AZLU) led on the LLM Judge (41.53) despite a lower Cosine score (59.15), highlighting a divergence between metric-based and qualitative evaluation.

5.2 Results of the Human Evaluation

The captions were evaluated by a human based on four main criteria: accuracy, completeness, conciseness, and cultural relevance.

Evaluation by native speakers yielded strong scores across key qualitative metrics: Cultural Relevance (3.20), demonstrating effective conveyance of cultural context; Conciseness (3.44), indicating direct and succinct phrasing; Accuracy (3.16), confirming factual alignment with image content. The lower Completeness score (2.88) suggests occasional omissions of finer contextual details.

5.3 Analysis: The Impact of Design Choices

Our captioning system prioritized the exclusive use of Modern Standard Arabic (MSA) to ensure linguistic coherence and prevent dialectal variation. Emphasis was placed on achieving succinctness while preserving cultural and historical accuracy. This methodology generated accurate and concise captions, with strengths in both conciseness and accuracy contributing to strong overall performance.

Participants	Cultural Relevance	Conciseness	Completeness	Accuracy
BZU	3.24	2.76	3.08	2.92
AZLU	3.20	3.44	2.88	3.16

Table 2: Human evaluation results for Subtask 1.

Human evaluation (Table 2) reveals a trade-off: BZU excels in Cultural Relevance and Completeness, while AZLU scores higher in Concise and Accuracy, suggesting a contrast between contextual nuance and precise succinctness.

5.4 Error Analysis: System Mistakes, Confusion Matrices, Error Types

Despite strong overall performance, several areas for improvement were identified:

Linguistic Errors: Occasional use of non-standard MSA terms due to dialectal interference.

Cultural Errors: Due to the absence of location data in some images and limited knowledge of specific locales, unidentified places are designated as 'Palestine'.

Visual Understanding Errors: Challenges in interpreting fine-grained, culturally nuanced scene details.

The evaluation was conducted on the whole dataset after collecting the whole captions in one csv file containing the batch id,image id, and the written caption.

5.5 Distinction between Official vs. Post-Submission Results

Based on the preliminary assessment, our team (AZLU) performed well, particularly in Conciseness and Accuracy, demonstrating the capacity to generate understandable and culturally relevant captions.

6 Limitations

While the Dataset addresses a significant gap in Arabic image captioning resources, it possesses a limitation that presents an opportunity for future work. The dataset's scale, with 500 image-caption pairs, is sufficient for initial benchmarking but remains limited for training large-scale models from scratch without significant data augmentation or transfer learning. A larger-scale dataset would be necessary to achieve state-of-the-art performance and improve model generalization.

7 Conclusion

This paper detailed a contribution to the ImageEval 2025 Shared Task: a manually annotated dataset of Arabic image captions in Modern Standard Arabic (MSA). By enforcing strict linguistic guidelines and prioritizing cultural relevance, we addressed key challenges in Arabic captioning, such as dialectal variation and a lack of public datasets. Our results demonstrated strong performance in conciseness and accuracy, validating the annotation methodology.

While some errors in cultural disambiguation and completeness were observed, this dataset provides a foundational resource. Future enhancements could include expanding the dataset size, integrating multimodal pre-training, and leveraging domain-specific lexicons. This work aims to advance Arabic natural language generation and foster greater inclusion of underrepresented languages in global research.

Acknowledgments

The authors thank the organizers of the ImageEval 2025 shared task, Mustafa Jarrar, Ahlam Bashiti, and Alaa Aljabari, for their work and guidance. We also acknowledge the image annotators.

References

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.

Jonathan Emami, Pierre Nugues, Ashraf Elnagar, and Imad Afyouni. 2022. Arabic image captioning using pre-training of deep bidirectional transformers. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 40–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Living Arabic Project. n.d. Living arabic project. https://www.livingarabic.com/.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for Arabic image captioning with gemini decoder. In *Proceedings of ArabicNLP 2023*, pages 1–11, Singapore (Hybrid). Association for Computational Linguistics.

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