# BZU-AUM at ImageEval 2025 Shared Task: An Arabic Image Captioning Dataset for Conflict Narratives with Human Annotation

**Mohammed AlKhanafseh**
Birzeit University
malkhanafseh@birzeit.edu

**Ola Surakhi**
American University of Madaba
o.surakhi@aum.edu.jo

**Abdallah Abedaljalill**
Birzeit University
1212725@birzeit.edu

## Abstract

This paper presents a new Arabic image captioning dataset created for the ImageEval 2025 Shared Task. The dataset focuses on images related to conflict, resistance, and everyday life under occupation. Each image is paired with a Modern Standard Arabic caption of 40–70 words that describes what is shown and adds cultural or emotional context. To help annotators write rich and consistent captions, we used prompt-based guidelines, including step-by-step reasoning and writing from specific roles such as journalists or humanitarian observers. This method produced captions that are both descriptive and meaningful. The dataset fills an important gap in Arabic resources, especially for sensitive and historically significant topics. It can be used to train and evaluate Arabic vision language models, test multilingual AI systems, and support applications in journalism, education, and cultural preservation.

## 1 Introduction

The ImageEval 2025 Shared Task encourages multilingual, culturally sensitive image captioning by having participants create Arabic captions for pictures that call for emotional, historical, or cultural knowledge. We make a contribution by compiling a dataset of Arabic captions that have been human-annotated for photos that show resistance, conflict, and day-to-day life under occupation. A crucial task in computer vision and natural language processing, image captioning produces natural language descriptions for visual content and facilitates uses like content indexing, accessibility, and visual storytelling. Even though the field has been advanced by large datasets like MS-COCO citelin2014microsoft, Flickr30k citeyou2016image, and Visual Genome citekrishna2017visual, they limit culturally rich or emotionally charged narratives by concentrating on English and generic domains. Despite being widely spoken, Arabic is still under-represented, which limits the use of large language models and vision-language in these situations. With thorough, human-written captions in formal Arabic (40–70 words) that capture both the visible content and any underlying cultural or emotional meaning, our dataset fills this gap. We used Chain-of-Thought (Wei et al., 2022; Kharma et al., 2025) and Role-Based prompting (Bubeck et al., 2023) to guarantee consistency and depth, directing annotators to reason methodically and adopt viewpoints similar to those of a journalist or witness. This tool facilitates the development of socially conscious AI systems, the training and assessment of Arabic vision-language models, and the improvement of multilingual LLMs. Additionally, it offers a standard for the ImageEval 2025 Shared Task, allowing models to be assessed on culturally relevant and contextually rich captions.

In the remaining portion of the paper, relevant work, dataset construction, annotation methodology, caption analysis, use cases, and future directions are discussed.

## 2 System Overview

In this section, we describe the dataset creation and annotation process used in our submission to the ImageEval 2025 Shared Task.

### 2.1 Dataset Composition and Historical Context

With an emphasis on the Palestinian experience as it has been shaped by resistance, displacement, and colonisation, this dataset tackles the dearth of culturally rich and emotionally charged Arabic image captioning data. In contrast to other datasets that contain neutral, apolitical content, it contains images of life under occupation, acts of protest and survival, and aspects of cultural continuity and defiance. These images function as digital repositories of memory and identity in addition to being training

*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 395–400
November 8-9, 2025 ©2025 Association for Computational Linguistics

data for vision-language models.

Two datasets of 250 images each, arranged into five thematically distinct sheets, were selected for analysis. As outlined in Table 1, each sheet focusses on a distinct facet of Palestinian history and daily life, guaranteeing coverage of both traumatic and resilient experiences.

Our dataset's imagery captures the nuanced historical and sociopolitical background of Palestine. Over 700,000 Palestinians were displaced and over 400 villages were destroyed during the 1948 Nakba, which was the result of political repression, house demolitions, military raids, and growing Zionist settlement during the British Mandate (1917–1948). Images of destroyed homes, military checkpoints, refugee camps, civilian resistance, and everyday resiliency were produced by later events, such as the 1967 Six-Day War and the ongoing occupation of the West Bank, Gaza, and East Jerusalem. This range of adversity and resilience is reflected in the photographs we have chosen. Archival collections, documentary photography, and publicly accessible materials that adhere to ethical and legal guidelines are examples of sources. While avoiding exploitative or dehumanising content, each image was carefully assessed for its emotional and historical significance. The captions emphasise social cohesion, cultural pride, and dignity while highlighting both suffering and resiliency.

In addition to offering top-notch, ethically sourced content for training and assessing Arabic vision-language models, this curation approach guarantees that the dataset portrays a complex, multi-layered Palestinian narrative that is frequently missing from widely used computer vision datasets.

### 2.2 Prompt-Guided Annotation Strategy

We used a structured prompt engineering approach to generate high-quality captions for the shared task, allowing for expressive captions that go beyond straightforward image descriptions. Formal captions of 40–70 words per image were written by native Arabic speakers who had received training in descriptive writing and sociopolitical context.

There are two primary methods for guided annotation: Role-Based prompting (Bubeck et al., 2023) and Chain-of-Thought prompting (Wei et al., 2022; Kharma et al., 2025). Methodical reasoning was promoted by Chain-of-Thought prompting, in which annotators first discussed observable elements (people, objects, and setting), then thought about actions or events, and lastly discussed historical, symbolic, or emotional ramifications. A child standing next to debris, for instance, could be explained not only in terms of the obvious damage but also in terms of the larger context of displacement or societal memory. As a result, the captions were contextually rich and semantically layered. Annotators were given distinct viewpoints, such as journalist, eyewitness, or humanitarian, along with suggested questions and tones for each role, thanks to role-based prompting. This method maintained thematic coherence while expanding the emotional and stylistic scope of captions.

Grammar, clarity, cultural correctness, and the absence of bias or conjecture were all examined in each caption. When an image had more than one caption, narrative and emotional significance were given priority during the selection process.

This dataset creates insightful, culturally relevant captions and allows for deeper interaction with images through the integration of structured prompting. It offers a standard for assessing Arabic vision-language models on content that demands both factual accuracy and narrative depth, and it methodologically advances socially conscious AI.

### 2.3 Experimental Setup

**Data.** We use 500 images with one caption each, organized in two batches (250/250). Each row has image_id, caption, and batch_id. Text is UTF-8 and in Modern Standard Arabic (MSA).

**Annotation.** Native Arabic speakers wrote the captions in a spreadsheet interface. They followed the template in Section 3.2 (role = journalist, eyewitness, or humanitarian). Rules: 40–70 words, MSA only, no speculation, no identification of minors.

**Review.** We used a two-pass review. Checks covered length, grammar, MSA register, factual grounding, tone, and role. Items that failed were corrected or replaced.

**Stats.** The numbers in Section 4 were computed with a simple script: word count by whitespace tokens and sentence count by Arabic/English punctuation.

**Packaging.** We provide a CSV with image_id, caption, batch_id. Prompt texts and scripts will be released after review.

## 3 Results and Analysis

This section reports the caption characteristics used in our ImageEval 2025 shared-task submission.

### 3.1 Linguistic and Structural Characteristics of the Captions

For ease of clarity, formality, and cultural depth, all of the captions are written in Modern Standard Arabic (MSA); the use of colloquial language was avoided given that MSA is preferred for academic, journalistic, and historical concerns. All captions are around 50 words in length, but could range from 15 words to 100 words in length. This is enough room to describe the visual content, while also adding context that was cultural, historical, and emotional. Captions are often more than facts, they represent a thoughtful and interwoven narrative that conveys historical significance and meaning.

Using formal, understandable language, this caption opens with an objective, detailed description of the visual scene, emphasising the subject's posture, surroundings, and companions. It then deciphers traditional attire and facial expressions as symbolic expressions of resistance and identity. Lastly, it highlights the image's political and cultural significance by placing it within a larger historical narrative. Reflecting the Chain-of-Thought prompting technique employed during annotation, the structure logically moves from particular visual details to general historical significance.

Similarly, foe example this caption presents a historic landmark with rich cultural connotations:

"تظهر هذه الصورة التاريخية قبة الصخرة المشرفة في القدس، بقبتها الذهبية المميزة وعمارتها الإسلامية الأصيلة، محاطة بأشجار السرو الشامخة والساحات الرحبة للحرم الشريف. تمثل هذه اللقطة النادرة من أوائل القرن العشرين جمال المسجد الأقصى وقداسته، حيث يقف هذا المعلم الإسلامي شاهداً على تاريخ فلسطين العريق وحضارتها الإسلامية. تجسد الصورة الهدوء والسكينة التي تخيم على هذا المكان المقدس، الذي يحمل في طياته ذكريات الإسراء والمعراج ومكانة القدس الخاصة في قلوب المسلمين في جميع أنحاء العالم."

Here, the caption highlights the landmark's religious and cultural significance while creatively describing the surrounding landscape and architectural beauty. Words like "القداسة" and "السكينه", arouse the reader's emotions and spirituality, urging them to value the location beyond its outward appearance. The caption maintains coherence and narrative flow in accordance with the dataset's guidelines by fusing together visual, historical, and cultural layers.

A third example depicts a historically significant landscape:

"تُظهر هذه الصورة التاريخية النادرة جبل الزيتون من باب النبي داوود في القدس، حيث تمتد التلال المقدسة والوديان التي شهدت أحداثاً تاريخية ودينية عظيمة عبر القرون. يبدو في المشهد جزء من أسوار القدس القديمة الحجرية والمباني التراثية المتناثرة على سفوح الجبل، بينما تظهر في الأفق مئذنة تشير إلى الحضور الإسلامي العريق في المدينة المقدسة. تحكي هذه اللقطة من أوائل القرن العشرين قصة القدس الخالدة، بتضاريسها الوعرة وتاريخها المتجذر في قلب فلسطين، مجسدة الطابع الروحاني والثقافي الذي يميز هذه الأرض المباركة."

This caption incorporates detailed topographical description, historical references, and spiritual symbolism, capturing the multifaceted nature of the depicted scene. The use of descriptive phrases like التلال المقدسة and ""التاريخ المتجذر reflects an elevated linguistic style that enriches the viewer's understanding. The caption also conveys a temporal dimension by situating the image historically "من أوائل القرن العشرين", enhancing its documentary value.

All of the captions in this dataset have a similar structure, in that they describe the visible components in the picture first, and then provide an interpretation of the image based on a historical, cultural, and/or emotional context, in order to introduce meaning to the submissions to train vision-language models. Role-based prompting provides a way to translate the different perspectives: witness or humanitarian perspectives use emotional, personal language, whereas journalistic captions sought objectivity and clarity. Having multiple voices is important to allow for multiple perspectives of conflict, identity, and cultural heritage to contribute to a cohesive narrative. The dataset with the relatively balanced length of captions, formality of Modern Standard Arabic, and content that is culturally meaningful offers the opportunity to contribute to the understanding of Arabic vision-language models in sensitive and complex domains.

### 3.2 Dataset-level

Table 3 shows the dataset is length-controlled yet varied: mean ≈54 words (median 51), 79.2% within the 40–70 target, and about 2.7 sentences per caption.

### 3.3 Shared Task Evaluation Results

With a BLEU-4 score of 0.41 and a ROUGE-L score of 0.56, our system demonstrated a high degree of phrasing and structural alignment with reference captions. Additionally, it received an LLM judge score of 32.42 and a cosine similarity mean of 65.53, which indicate linguistic fluency and semantic closeness. According to these findings, our system outperformed all others in terms of meaning accuracy and fluency. Only a few minor visual details were left out of the captions, which did a good job of capturing the cultural and historical context.

## 4 Use Cases and Future Directions

This Arabic image captioning dataset, emphasizing cultural identity, resistance, and conflict, enables research on vision-language models and the development of socially significant AI.

### 4.1 Use Cases

This dataset pairs historically significant images with culturally rich captions to support a variety of important applications:

- **Multimodal LLM and Arabic VLM training:** Improves models to comprehend intricate historical, cultural, and affective image contexts.
- **Assistive technologies:** Enhances accessibility for Arabic speakers by offering more detailed, contextualised descriptions of images..
- **Cultural heritage preservation:** Aids in recording and disseminating Palestinian history and regional conflicts to educational institutions.
- **Digital archiving:** Makes it possible to create searchable, semantically rich archives that aid in research and preserve collective memory.
- **Journalism and humanitarian work:** Automates fact-checking of photos from conflict areas and sensitive, accurate storytelling.

### 4.2 Future Directions

Although this dataset provides a useful tool for researching the relationship between language, culture, and history, much more could be done to expand its use, boost its influence, and guarantee responsible use. The following areas will be the focus of future efforts:

- Increase the number and variety of images.
- Incorporate more languages and a greater variety of Arabic dialects.
- Examine how language models handle bias, cultural quirks, and emotions using the dataset.

## 5 Limitations and Ethical Considerations

Even though this dataset has special historical and cultural significance, responsible use requires acknowledging its limitations and ethical issues. These elements are summed up in the points that follow:

**Limitations:**

- **Small size:** Only 250 images, limiting diversity of visuals and contexts (Torralba and Efros, 2011).
- **Language:** Captions in Modern Standard Arabic (MSA) ensure consistency but exclude dialectal nuances (Abdul-Mageed et al., 2023).
- **Generalizability:** Integration with other datasets is recommended (Nguyen and Ploeger, 2025).

**Ethical Considerations:**

- Includes delicate and possibly upsetting material (conflict, trauma, and displacement).
- Risks of abuse, deception, or retraumatization.
- Guidelines for annotations placed a strong emphasis on factual, courteous descriptions that shunned bias or sensationalism.

## 6 Conclusion

The lived experiences of Arabic-speaking communities impacted by historical trauma and conflict are being connected to AI for the first time with this dataset. It does more than just describe pictures; it tells stories with social and emotional significance, highlighting cultural heritage, resiliency, and resistance. AI can now interpret images while honouring the voices and histories they represent thanks to this method. To guarantee that future AI tools continue to be morally and culturally appropriate, we encourage continued cooperation between linguists, historians, AI researchers, and local communities. The AI community can support social justice and cultural memory preservation by growing and improving these datasets and encouraging inclusive practices.

In the end, this resource shows how AI can be used not only for Arabic vision-language research but also as a tool for empathy, comprehension, and historical preservation, encouraging work that respects human experience and crosses cultural divides.

## Appendix: Additional Tables

Table 1: Overview of key historical events and figures in Palestinian and Rif resistance history.

| Topic | Details |
|---|---|
| Mohammed bin Abdelkrim El Khattabi El Ouryaghli | Moroccan judge and fighter, leader of the Rif resistance against Spanish and French colonialism, founder of the Republic of the Rif (1882–1963). |
| 1948 War | Started May 15, 1948; Arab-Zionist conflict caused mass Palestinian displacement ("Nakba"). |
| Jerusalem | Capital of Palestine, historic and religious city with Al-Aqsa Mosque; strategic central highlands location. |
| Palestinian Cities Occupied in 1948 | Haifa, Acre, Jaffa: cultural and commercial hubs; Nablus and Bethlehem: religious and historical significance. |
| British and Zionist Colonialism (1917–1948) | Palestinians faced repression and displacement during the British Mandate, resisted through uprisings, strikes, and armed struggle. |
| Zionist Attacks on Beirut | Israeli invasion caused widespread destruction and thousands of civilian casualties. |
| Lebanese Civil War (1975–1990) | Fifteen-year multifaceted conflict with 120,000 deaths and millions displaced, involving multiple sectarian, Palestinian, Israeli, Syrian, and international actors. |

Table 2: Summary of key historical events and cultural aspects of Palestine and the Rif region (second dataset).

| Topic | Details |
|---|---|
| British and Zionist Colonialism (1917–1948) | Persecution escalated under the British Mandate, including home demolitions, military raids, and displacement. |
| 1967 War | 1967 war caused Israeli occupation of key territories and further Palestinian displacement. |
| Jerusalem | Capital of Palestine; historic city with Al-Aqsa Mosque and Dome of the Rock, Islam's third holiest site |
| Palestinian Cities Occupied in 1948 | Haifa, Acre, Jaffa: historic trade centers; Nablus, Bethlehem: key religious and cultural sites. |
| Daily Life in Palestine | Markets, religious centers, rural herding, fishing, and glassmaking reflect social and economic diversity. |

Table 3: Caption statistics (overall).

| Statistic | Value |
|---|---|
| Images | 500 |
| Captions | 500 |
| Target length (words) | 40–70 |
| Mean words per caption | 53.99 |
| Median words per caption | 51 |
| Sentences per caption (avg.) | 2.72 |
| % within 40–70 words | 79.2% |
| Min / Max words | 3 / 148 |

## 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. *arXiv preprint arXiv:2310.16117*.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. 2023. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mohammed Kharma, Soohyeon Choi, Mohammed AlKhanafseh, and David Mohaisen. 2025. Security and quality in llm-generated code: A multi-language, multi-model analysis. *arXiv preprint arXiv:2502.01853*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Dong Nguyen and Esther Ploeger. 2025. We need to measure data diversity in nlp–better and broader. *arXiv preprint arXiv:2505.20264*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. 2023. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. 2023. Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2617–2625.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.