JEEM: Vision-Language Understanding in Four Arabic Dialects

Anonymous ACL submission

Abstract

We introduce JEEM, a benchmark designed to evaluate Vision-Language Models (VLMs) on visual understanding across four Arabic-speaking countries: **J**ordan, The **E**mirates, **E**gypt, and **M**orocco.[1] JEEM includes the tasks of image captioning and visual question answering, and features culturally rich and regionally diverse content. This dataset aims to assess the ability of VLMs to generalize across dialects and accurately interpret cultural elements in visual contexts. In an evaluation of five prominent open-source Arabic VLMs and GPT-4o, we find that the Arabic VLMs consistently underperform, struggling with both visual understanding and dialect-specific generation. While GPT-4o ranks best in this comparison, the model's linguistic competence varies across dialects, and its visual understanding capabilities lag behind. This underscores the need for more inclusive models and the value of culturally-diverse evaluation paradigms.

1 Introduction

Vision-language models (VLMs) have recently achieved notable improvements in tasks such as image captioning (IC) and visual question answering (VQA), benefiting from large multimodal training datasets and parameter scaling (LlamaTeam, 2024; Beyer et al., 2024; OpenAI, 2024). However, these models often struggle to generalize across culturally diverse and dialect-rich environments due to the over-representation of specific geographic regions (De Vries et al., 2019; Gustafson et al., 2023) and standardized language varieties in their training datasets (Pouget et al., 2024). Similarly, existing evaluation datasets predominantly feature Western-centric images and English text (Liu et al., 2021; Wang et al., 2024), while their non-English counterparts are often derived from the former, either through translation or relabeling of the same images (Changpinyo et al., 2023). This results in biased evaluation, which conceals the suboptimal performance of VLMs in geographically and dialectally diverse settings (Bhatia et al., 2024).

Figure 1: A sample from JEEM (Moroccan set). For brevity, only 2 Q&A pairs are shown.

Recognizing this gap, recent work has focused on the creation of culturally diverse multilingual VQA benchmarks, incorporating images and questions from various countries and languages (Liu et al., 2021; Pfeiffer et al., 2022; Changpinyo et al., 2023, *inter alia*). Among these, Arabic is rarely included, and when it is, it appears either in its standardized form (Modern Standard Arabic) (Tang et al., 2024) or a single dialect, such as Egyptian (Romero et al., 2024). This approach overlooks the cultural and dialectal diversity found among the ∼400 million speakers of this language.

[1] Our data is available at anonymous.

Arabic is an official language in 25 countries across North Africa and the Middle East. Despite the shared language, each country has a different history, geography, and consequently culture. These differences manifest in the objects, locations, and activities that visually characterize each region, as well as the lexical terms and implicit meanings associated with them. For example, the traditional clothing item in the Gulf, the 'kandura' (a long white robe worn by men) differs subtly from the 'djellaba' worn in Upper Egypt, each reflecting regional identity and invoking different societal norms. On a linguistic level, differences are found not only in terms of lexicon, but also in phonetics and syntax, sometimes making mutual intelligibility challenging even among native Arabic speakers.

To address the challenges posed by the cultural and dialectal diversity of Arabic, we introduce JEEM, a benchmark dataset spanning one representative dialect from each dialectal region (Habash, 2010): **J**ordanian (Leventine), **E**gyptian, **E**mirati (Khaleeji), and **M**oroccan (Maghrebi). JEEM comprises two core tasks: image captioning and visual question answering. These tasks enable the evaluation of VLMs in terms of their ability to recognize and appropriately reason about cultural elements, such as traditional clothing, local artifacts, and social settings, while utilizing dialectal language.

We benchmark five VLMs on JEEM and measure performance in terms of standard count-based metrics, GPT4-based evaluation, and human evaluation. This comprehensive evaluation protocol allows us to reliably compare the five VLMs, identifying performance gaps in all models, including the top-ranking one, GPT-4o. We also evaluate the automatic metrics against human judgments and provide recommendations for future evaluations on JEEM and other Arabic benchmark datasets.

2 Related Work

2.1 Why Culture Matters

Prior studies reported performance disparities across cultures on machine learning tasks such as object recognition (De Vries et al., 2019; Gustafson et al., 2023), geolocalization (Pouget et al., 2024), mutlimodal retrieval (Kádár et al., 2018; Buettner and Kovashka, 2024) and visual question-answering (Romero et al., 2024). These disparities are commonly attributed to biases in the data on which models are trained, which tends to over-represent high-income geographic regions, and in particular Western ones (De Vries et al., 2019; Gustafson et al., 2023; Pouget et al., 2024).

People from different cultures use different objects, have different traditions, and occupy different physical environments, resulting in different visual experiences and associations. Culture also affects perception and language: it determines whether a more general or a more specific term will be used to refer to an object, how the importance of background objects will be ranked with respect to foreground objects, and what objects will be mentioned in a caption or omitted (Nisbett and Masuda, 2013; Buettner and Kovashka, 2024).

2.2 Vision-Language Resources in Arabic

Image Captioning Early work on Arabic image captioning (Jindal, 2017; Mualla and Alkheir, 2018) relied on the machine translation of existing datasets (primarily MSCOCO (Lin et al., 2014) and Flickr8K (Hodosh et al., 2013)), sometimes including human validation (ElJundi et al., 2020) or human translation for a subset of the data (Almuzaini et al., 2018). AraCOCO (Mohamed et al., 2023) features 500 images from the MS COCO test set, captioned by Arabic speakers. While the annotators often mentioned details that did not appear in the original English caption, attesting to the difference in cultural perspectives, the captioned images were not sourced from the Arab-speaking world. Moreover, these captions are in MSA and follow the same short simplistic format found in COCO, lacking in dialectal and cultural understanding.

Visual Question-Answering While several multilingual datasets focus on culturally relevant VQA, many exclude Arabic (Gao et al., 2015; Gupta et al., 2020; Pfeiffer et al., 2022; Changpinyo et al., 2023). VAQA (Kamel et al., 2023) relabels MS COCO images in MSA, again limiting the cultural relevance and dialectal coverage of the data. Some works in text-only QA address this issue through manual cultural alignment (Alyafeai et al., 2024) or sourcing data from Arab countries directly (Koto et al., 2024), but lack the visual component. Others incorporate images but are limited to a single Arabic variety (Tang et al., 2024; Romero et al., 2024), or rely on synthetic questions that are not always visually grounded (Alwajih et al., 2024).

3 Dataset Construction

JEEM consists of images originating from four Arabic-speaking countries covering four distinct

dialectal regions: Jordan (Levantine), Emirates (Gulf), Egypt (Egyptian), and Morocco (Maghrebi). Each image is annotated by native speakers of the target dialect with image captions in both MSA and dialect, and question-answer pairs in dialect.

Team Organization and Recruitment The annotation process was led by four native speakers of the target dialects, each with a background in computational linguistics or natural language processing, hereafter referred to as team leaders.

The annotator recruitment process began with a free qualification task designed to identify annotators who met the following criteria: *i*) had relevant professional experience; *ii*) were native speakers of the target dialects; *iii*) could produce high-quality image captions. As part of the qualification task, candidates wrote a caption for one image in both the target dialect and MSA. Each submission was carefully reviewed by a team leader. The candidates who performed best in terms of fluency and relevance were subsequently invited to join the project. This process led to the recruitment of 10, 8, 10, and 9 annotators for Jordan, the Emirates, Egypt, and Morocco, respectively. [2] Their sociodemographic statistics, collected through a voluntary survey, can be found in Appendix A.

Annotation Setup The data collection process is based on how a visually impaired user might interact with a smart assistant: given an image with which the user wishes to engage (Step 1), the smart assistant would offer an initial description of the image (Step 2); at this point, the user might ask clarifying questions and inquire about further details (Step 3), to which the assistant would provide an answer (Step 4). We do not claim this procedure to accurately represent the experience and needs of visually impaired users, but it serves as a useful framework for guiding annotators on how to engage with the task, and for collecting natural questions born out of a genuine information scarcity. The process is visualized in Figure 7 in Appendix B.

Step 1: Image Collection The objective of this step is to gather diverse, publicly available images that represent typical daily life in the target regions. To this end, we collected images from three sources: *i*) Wikimedia archive, where images were sampled from categories under the tag `Category:<country>_by_topic` (all subject to a Creative Commons license). *ii*) Flickr archive under a Creative Commons license: the images were retrieved using tags such as country names, city names, and names of important places. *iii*) Personal archive: coauthors of this paper and team leaders contributed images from their personal collections that show typical scenes of daily life in their region of origin. They also reviewed and filtered all images sourced from Wikipedia and Flickr to ensure appropriate and informative selection. Refer to Table 1 for final image counts in JEEM.

[2] Throughout the project, we observed significantly less involvement from the Emirati annotators compared others.

Country	Image Count	Average Length		Unique Words	
		DA	MSA	DA	MSA
Jordan	606	46	52	8,933	9,751
Emirates	150	41	44	2,453	2,574
Egypt	863	58	63	10,700	12,941
Morocco	577	52	52	7,822	8,161

Table 1: Dataset statistics: image count, average caption length, number of unique words in the JEEM dataset.

Step 2: Image Captioning The task is to write a description of the given image in both MSA and dialect. Annotators were instructed to write in their dialect first to encourage spontaneous writing. They were instructed to provide descriptions that are detailed enough to convey the content to someone who cannot see the image, including details specific to their region.

Step 3: Question Writing The task is to write five questions in dialect based on the given image description (the image is not shown to the annotator). The questions should be independent of each other and aim at a better understanding of what is happening in the unseen image. To avoid repeated exposure to images, annotators assigned to write a caption for a particular image were not assigned to write questions for the same image.

Step 4: Question Answering The task is to answer five questions in dialect, based on the corresponding image and captions. If it is not possible to answer a question (e.g., the image does not contain the necessary information), annotators were instructed to indicate that the image lacks sufficient information. Answers should be based on the image and a general understanding of its context.

Annotators in each dialect group were assigned to annotate images specific to their region. In addition, a set of 25 images per dialect was manually selected to form a *shared* pool of culturally distinctive places, dishes and objects. These images were

annotated by all four regional teams to enable the exploration of cross-cultural perspectives.

Task Review Each submitted task was reviewed by the respective team leader. Reviewers could reject a task and reassign it to another annotator, edit and accept the task, or accept it as is. Additionally, reviewers were allowed to skip a task if the image or the writing appeared inappropriate or irrelevant. The team leaders collaborated closely with the annotators throughout the annotation project, providing suggestions for improvement and exchanging feedback in a group chat.

Dialect Diversity Each annotator could complete a limited number of tasks per day to avoid having a small number of annotators dominating the annotations. See Appendix A for the distribution of tasks completed by each annotator. To ensure dialect diversity, annotators were encouraged to use the most natural language for their local area. If they encountered an unfamiliar word or phrase in writing from previous steps, they were instructed to ask in the group chat for clarification.

The annotation was carried out on the `Anonymized` platform. Refer to Appendix A for time estimates associated with the `JEEM` collection. The detailed annotation guidelines made available to the annotators can be found in Appendix B.

4 Data Analysis

4.1 Data Distribution

`JEEM` consists of 2,196 annotated images, distributed across the four dialects, as shown in Table 1. The largest portion of images belongs to Egypt, followed by Jordan and Morocco, while the Emirates is represented in 7% of the data. The table also reports statistics on the image captioning part of the dataset, including average caption length and the number of unique words used in each dialect.

Image Topics The images are organized into 13 thematic categories, including places, events, arts, nature, education, transport, food, trade, technology, characters, and games. These categories were identified after building the dataset, where we prompted GPT-4o mini (gpt-4o-mini-2024-07-18) (OpenAI, 2024) in multiple rounds: first, to assign a topic to each MSA caption, and then to group the identified topics into the final categories (see the prompt in Appendix E, Figure 13). A detailed breakdown of subtopics within each category is provided in Table 6 in Appendix C, while

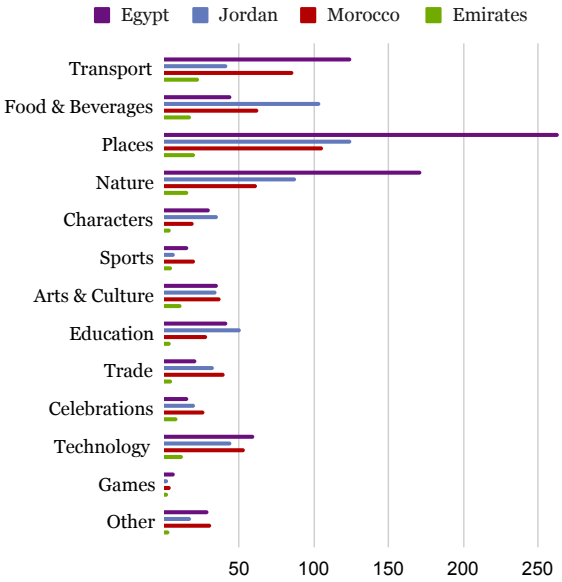


Figure 2: Topic distribution per dialect.

Figure 2 visualizes the topic distribution across dialects. Places emerge as the most common topic across images from all countries. However, the distribution of other prominent topics varies by region: nature was the second most frequent topic in Egypt, while food and beverages (F&B) dominated in Jordan, and transport was prominent in both the Emirates and Morocco.

Image Captions In written form, MSA and dialectal Arabic exhibit distinct variations in morphemes, sentence structure, and spelling conventions, which serve as indicators of dialectal influence in text (Keleg et al., 2023). This is evident in the variation of the average number of words used in captions across different dialects, as shown in Table 1. Emiratis tend to use the fewest number of words in their captions, averaging 41 words per caption. In contrast, Egyptians write significantly longer captions, averaging 58 words per image.

Questions and Answers The total number of QA pairs across dialects is 10,890. In order to gain insight into the type of questions asked, we employed few-shot prompting of GPT-4o mini (gpt-4o-mini-2024-07-18). The prompt defines four distinct question types (Descriptive, Quantitative, Categorical, and Yes/No) and, provides a detailed explanation of its defining characteristics with three examples in different dialects (see the prompt in the Appendix, Figure 12). The distribution of question types across dialects is shown in Table 2, alongside some examples. The most prevalent type of questions is Descriptive, accounting for 45.92% of

Type: Descriptive **Example:** What are the people on the roof wearing?	**Percentage:** 45.92 الناس اللي على السقف لابسين إيه؟
Type: Categorical **Example:** Are these people in the kitchen men or women?	**Percentage:** 18.83 هاد الناس اللي فكوزينة واش رجال و لا عيالات؟
Type: Quantitative **Example** How many boats can we see in the picture?	**Percentage:** 8.83 كم طراد نقدر نشوفه في الصورة؟
Type: Yes/No **Example** Does it look like they're cooking something on the stove?	**Percentage:** 26.42 مبين انهم حاطين اشي عالنار؟

Table 2: Question type distribution across JEEM.

Jordanian
الكراوية او الدبس
karawya or dibs

Emirati
حلوى عمانية
Omani halwa

Egyptian
بودنج شيكولاتة
Chocolate pudding

Moroccan
شكلاط
Chocolate

Figure 3: Image of a Omani Halwa (image sourced from the Emirati set) shared with all annotators. The non-Emirati captions demonstrate an incorrect identification of the dessert. Refer to Figure 8 for full captions.

the total, followed by Yes/No questions at 26.42%, Categorical questions at 18.83%, and Quantitative questions at 8.83%.

4.2 Cultural Aspects

We manually explored the shared pool of 100 images captioned in all four dialects to gain an understanding of how cultural perspective shapes perception. One notable example is shown in Figure 3. It involves an image of Omani Halwa, a traditional Gulf dessert made from margarine, sugar, rose water, and semolina. As illustrated in the captions in the four dialects, only the Emirati annotator correctly identified it as Omani Halwa. In contrast, the Jordanian annotator misidentified it as Karawya, a visually similar dessert with slight variations in texture and color, and both the Moroccan and Egyptian annotators mistakenly described it as a chocolate dessert, showcasing the diverse regional influences on object recognition.

5 Benchmarking VLMs

We benchmark five recent Arabic-capable VLMs on JEEM: Maya (Alam et al., 2024), PALO (Rasheed et al., 2025), Peacock (Alwajih et al., 2024), AIN (Heakl et al., 2025), and AyaV (Cohere, 2025). All of these models were trained on Arabic data, sometimes authentic, but often translated from English. For completeness, we also evaluate GPT-4o (gpt-4o-2024-08-06) (OpenAI, 2024) on JEEM, as it has been shown to achieve strong performance on various Arabic tasks (Alyafeai et al., 2023). We carry out an extensive meta-evaluation of different natural language generation metrics on the image captioning task, and then apply the most successful metrics to the VQA task.

5.1 Image Captioning

5.1.1 Evaluation Metrics

We include four traditional captioning metrics: CIDEr (C) (Vedantam et al., 2015), ROUGE (R) (Lin, 2004), BLEU (B) (Papineni et al., 2002), and BERTScore (BSc) (Zhang et al., 2020). For BERTScore, we use CamelBERT (Inoue et al., 2021), which was trained on both MSA and dialectal Arabic. We report recall rather than F1 score, following Zhang et al. (2020). Given the morphological complexity and dialectal diversity of Arabic, we expect standard metrics to prove suboptimal.

We also conduct a **GPT-4-as-a-Judge** evaluation following Tong et al. (2024), where the LLM sees both the input image and the reference caption when assessing a generated caption. This setup enables judgments grounded both in image and reference text. We perform the evaluation according to four criteria, based on Liu et al. (2023): *i*) *Consistency (Con)* evaluates whether the caption matches what is shown in the image; *ii*) *Relevance (Rel)* evaluates whether the caption describes the most important elements in the image; *iii*) *Fluency (Flu)* evaluates how natural and fluent the text is; *iv*) *Dialect Authenticity (DAuth)* evaluates whether the caption represents the target dialect. Each criterion is evaluated on a five-point Likert scale, where 1 indicates failure to meet the criterion and 5, full compliance. We use GPT-4 for this evaluation (gpt-

	Model	Traditional Metrics				GPT-4-as-a-Judge*				DCScore*			ALDi	Human Eval*			
		B	C	R	BSc	Con	Rel	Flu	DAuth	Precision	Recall	F1-score		Con	Rel	Flu	DAuth
MSA	AIN	4.00	1.05	7.46	80.31	2.55	2.49	4.20	-	69.87	41.87	48.32	-	-	-	-	-
	AyaV	4.10	0.76	9.85	90.36	3.10	3.27	4.53	-	79.65	**64.31**	**70.43**	-	-	-	-	-
	Palo	4.26	1.76	9.48	**90.46**	2.48	2.56	4.02	-	64.73	42.32	50.08	-	-	-	-	-
	GPT-4o	**5.87**	**7.27**	**10.61**	90.35	**3.67**	**3.75**	**4.77**	-	**87.54**	56.89	68.00	-	-	-	-	-
JO	AIN	2.19	0.45	5.57	81.55	2.71	2.72	4.36	2.60	69.59	45.07	51.57	0.48	2.53	2.60	3.74	1.46
	AyaV	2.68	0.83	7.59	89.34	3.18	3.28	4.39	2.95	80.83	**65.07**	**71.49**	**20.93**	3.80	3.75	3.98	2.31
	Palo	2.05	0.68	6.63	**90.73**	2.79	2.77	4.35	2.63	68.54	47.43	54.97	0.00	2.93	2.89	4.00	1.27
	GPT-4o	**5.23**	**6.91**	**9.66**	90.72	**3.80**	**4.08**	**4.75**	**3.46**	**82.39**	56.75	66.39	19.57	4.05	4.05	4.35	3.27
AE	AIN	1.63	0.52	5.03	81.89	2.59	2.76	4.20	2.07	68.14	48.32	52.16	1.02	3.30	3.58	1.64	1.00
	AyaV	1.69	0.88	5.80	**90.24**	2.94	3.20	4.34	2.22	76.08	**66.34**	**70.15**	**4.82**	4.02	4.21	2.65	1.96
	Palo	1.50	0.29	5.75	89.43	2.53	2.71	4.16	1.88	65.90	46.92	53.28	0.00	2.38	2.78	1.16	1.00
	GPT-4o	**3.19**	**2.73**	**7.21**	89.03	**3.58**	**3.84**	**4.74**	**2.58**	**81.47**	57.62	66.53	3.42	3.24	3.32	2.20	2.20
EG	AIN	2.08	0.31	5.20	79.60	2.36	2.32	4.26	2.45	67.65	38.00	43.90	1.75	3.44	3.31	4.11	1.87
	AyaV	2.87	0.83	7.85	89.82	2.61	2.73	4.16	3.65	72.78	**56.18**	**62.41**	**54.24**	3.77	4.04	3.99	3.63
	Palo	2.05	0.52	6.37	91.10	2.22	2.47	4.15	2.42	63.15	38.96	47.09	0.00	3.52	3.79	4.59	1.33
	GPT-4o	**4.09**	**8.41**	**8.56**	90.64	**3.36**	**3.58**	**4.67**	**4.12**	**82.74**	50.83	62.06	49.86	4.43	4.41	4.65	4.45
MA	AIN	1.34	0.58	3.40	81.37	2.50	2.62	4.33	1.64	67.87	50.73	55.32	0.33	3.49	3.37	4.23	1.01
	AyaV	2.21	0.53	6.55	88.33	2.92	3.12	4.11	3.91	73.21	**62.44**	66.50	38.30	3.72	3.81	3.65	2.85
	Palo	1.06	0.46	3.76	89.47	2.82	2.93	4.30	1.79	70.58	50.95	58.33	0.00	3.93	3.98	4.72	1.00
	GPT-4o	**4.73**	**6.70**	**9.00**	**89.98**	**3.69**	**3.75**	**4.81**	**4.50**	**80.59**	57.79	**66.58**	**44.51**	4.65	4.65	4.50	4.41
	τ_c	19.79	11.78	15.69	10.62	39.56	31.54	10.88	47.27	36.93	40.16	40.63	40.11	-	-	-	-

Table 3: Traditional automatic metrics are calculated on the full dataset, whereas all starred entries (*) (GPT4o-as-a-Judge, DCScore, and human judgments) are computed on the same 350-image sample.

4-turbo-2024-04-09), a version intentionally distinct from the one used to generate the captions. Still, we are aware that bias from using GPT-4 to evaluate GPT-4o's generations can be a concern.

Therefore, we also include **DCScore** (Ye et al., 2025), a recently proposed evaluation method designed for detailed image captioning, which leverages the capabilities of GPT-4o in a structured, stepwise manner. DCScore decomposes the candidate and reference captions into primitive information units (PIUs): short, self-contained statements about objects, attributes, or relationships. DCScore is computed as an F1 score over the precision and recall of matched PIUs (see Appendix I for details.) Since the metric abstracts away from the surface form of the captions, we expect that bias from the choice of backbone LLM should be negligible.

Two complementary experiments, LLM-as-a-Judge evaluation without the image (reference-only condition) and evaluation using **HalFScore** (Chen et al., 2025), are included in Appendix G. Lastly, we use the **ALDi** (Keleg et al., 2023) model trained on the AOC-dataset to evaluate dialectness.

In order to establish which of the automatic metrics listed above are actually reliable and to extract further insights about model performance, we also conduct a **human evaluation**. The evaluation covers 100 images for Egyptian, Moroccan, and Jordanian (with three annotators per sample), and 50 images for Emirati (with one annotator per sample) due to limited availability of active annotators for this dialect. All model predictions, as well as the ground-truth captions, are evaluated according to the same four criteria as used in the GPT-4-as-a-Judge evaluation described above. In total, we obtained human evaluation judgements for 6,650 captions. This data will be made public to enable future meta-evaluation of automatic evaluation metrics for Arabic.[3] Following prior work (Hessel et al., 2021; Wada et al., 2024), we measure the quality of automatic evaluation metrics against human evaluation scores using **Kendall's Tau-c** (τ_c), which measures the correlation of ordinal data.

5.1.2 Results

Table 3 reports model performance across the five subsets of JEEM. The two weakest models, Maya and Peacock, are excluded here for brevity; their results can be seen in Table 10 in the Appendix H. The results for the ground-truth captions are also included in Table 8 in the Appendix F, and follow an expected pattern: scores are higher across the board compared to model-generated captions.

Metric Quality Traditional metrics like BLEU, CIDEr, ROUGE, and BERTScore show weak alignment with human judgments, with Kendall's Tau-c values ranging from 0.1 to 0.2. This reflects their limited suitability for Arabic. **GPT-4-as-a-Judge** improves correlation (up to 0.39) when given both

[3]See Appendix B for annotation guidelines and statistics.

the image and reference caption, but still falls short of modeling human preferences reliably. **DCScore** achieves the highest correlation (0.41), measured against the harmonic mean of human Consistency and Relevance scores. Its structured multi-step methodology makes it most suitable in this setting.

In terms of dialect authenticity, ALDi demonstrates strong agreement with human judgments. While GPT-4 achieves a higher overall Kendall's (τ_c) correlation (0.472 vs. 0.401), as in Table 3), both metrics have limitations in fully capturing human preferences. These findings underscore the broader challenge of designing automatic evaluation methods that reliably reflect human judgments across different Arabic dialects.

Model Comparison Based on human evaluation scores, we find that GPT-4o performs best across most dialects and criteria, particularly in Consistency, Relevance, and Fluency. However, it struggles with the Emirati dialect, particularly in Dialect Authenticity. This is likely due to underrepresentation of this dialect in its training data. Among the open-source models, AyaV consistently performs best. It achieves strong human scores in content alignment and reasonable scores in dialect authenticity in Jordanian, Egyptian and Moroccan. Palo matches AyaV in Moroccan and Egyptian in Consistency, Relevance and Fluency but its Dialect Authenticity is close to 1 (floor level). AIN tends to generate fluent captions but falls short in both visual grounding and dialectness.

Across all models, Fluency is the highest-scoring dimension, indicating that the LLMs can produce well-formed text. Meanwhile, Dialect Authenticity remains challenging, suggesting that models may be generating in MSA. We investigate this in §5.3.

5.2 Question Answering

5.2.1 Evaluation Metrics

We evaluate the VQA capabilities of four models, AyaV, AIN, Palo, and GPT-4o, which were the top performers in the image captioning task. We use GPT-4-as-a-Judge for evaluation, with the same four criteria as before. This metric is particularly suitable for descriptive questions, which constitute the majority of the JEEM VQA set (see Table 2.) Such questions are open-ended and admit multiple valid answers, rendering traditional string-matching metrics ineffective. While DCScore proved effective for image captions, its structured decomposition into primitive information units is not applicable to short-form VQA answers, which lack the richness and structure needed for that style of evaluation.

	Model	Con	Rel	Flu	DAuth
JO	AIN	2.41	2.55	4.11	3.04
	AyaV	2.76	2.96	4.22	2.55
	Palo	2.39	2.48	4.08	2.89
	GPT-4o	**3.56**	**3.70**	**4.67**	**4.26**
AE	AIN	2.78	2.83	4.19	2.93
	AyaV	3.00	3.02	4.26	2.57
	Palo	2.51	2.50	4.07	2.83
	GPT-4o	**3.64**	**3.72**	**4.56**	**3.87**
EG	AIN	2.18	2.26	3.72	2.92
	AyaV	2.63	2.72	4.10	2.66
	Palo	2.07	2.09	3.90	3.06
	GPT-4o	**3.26**	**3.36**	**4.58**	**4.54**
MA	AIN	2.22	2.36	3.68	2.37
	AyaV	2.60	2.78	4.06	2.10
	Palo	2.00	2.14	3.76	2.21
	GPT-4o	**3.52**	**3.71**	**4.58**	**4.56**

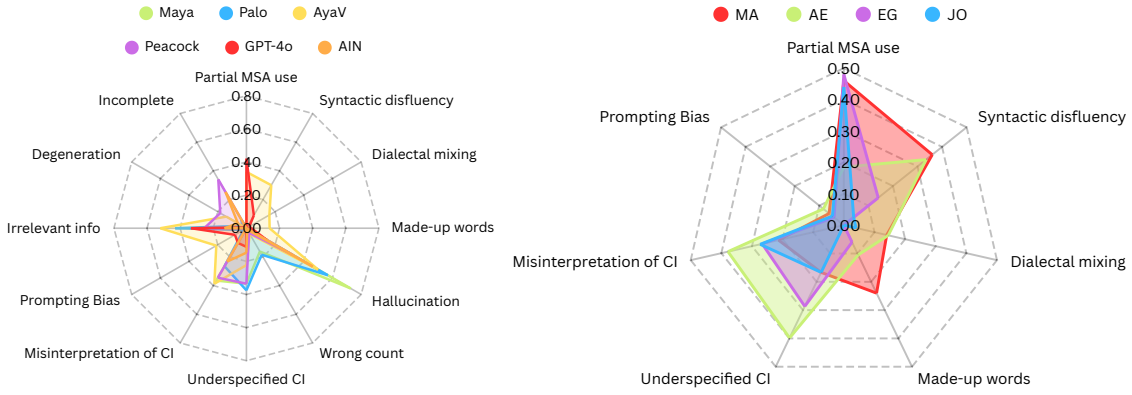
Table 4: GPT4-based evaluation of **question answering**, results for different Arabic varieties.

5.2.2 Results

The results are presented in Table 4. Again, we see that the fluency criterion yields the highest scores across all models and dialects. However, the semantic criteria of consistency and relevance lag well behind, especially for the Arabic VLMs. This suggests that the model-generated answers do not successfully address the user question. The dialectness of the Arabic VLMs is similarly low across most dialects. Meanwhile, GPT-4o scores best overall, despite some room for improvement on the semantic criteria, and reduced linguistic capabilities on the Emirati dialect. These findings align with the observations made in the image captioning task.

5.3 Error Analysis

To better understand the qualitative differences across models, we conduct a human analysis of error types, carried out by paper co-authors who are native speakers. We analyze a subset of 25 images–five from each country and five representing general Arab culture–with predictions from six models, each prompted in four dialects, resulting in a total of 600 evaluated predictions. We include the two models excluded from the main results (Maya and Peacock), to obtain clarity on the causes of their poor performance. Based on initial observation, we define a taxonomy of 11 error types that fall under the following categories: dialectal, visual, cultural, and generation. The types can be seen in Figure 4a and examples, in Table 11.



(a) Cross-Model (b) Cross-Dialect

Figure 4: Error type percentages averaged by dialect (a) and by model (b). For clarity, we omit the *complete MSA use* type, which is maxed out on all models except GPT-4o and Ayav. CI: cultural item.

Cross-Model Analysis We first compare model differences averaged across dialects (Figure 4a).

Since GPT-4o and AyaV are the only models with **dialectal** capabilities (see Table 3), we exclude the rest of the models from consideration under this category. While both GPT-4o and AyaV *partially use MSA*, AyaV shows a pronounced tendency of *making up words* (with an error rate of 14%) and *dialect mixing* (16%). Both models exhibit *syntactic disfluency* to varying degrees (9% for GPT-4o and 30% for AyaV.)

In terms of **visual** grounding, Maya stands out with the most *hallucinations*, with an error rate of 73%. Palo, AyaV, and AIN also exhibit notable *hallucination* tendencies. Maya and Palo additionally show *wrong object counts* and are overall least grounded in the visual modality.

In the **cultural** category, GPT-4o performs best with the lowest rates of *cultural underspecification* (11%) and *cultural misinterpretation* (10%). AyaV has the highest instance of *misinterpretation* (39%), while Palo tends to be overly vague (38%). AyaV can also suffer from *prompting bias* (21%).

Regarding **generation** quality, Peacock struggles with the highest rates of *degeneration* (18%) and *incompleteness* (34%). AIN also displays a relatively high rate of *incomplete* outputs (25%), a behavior not observed in other models. AyaV and Palo prominently introduce *irrelevant information* on 52% and 43% of their outputs, respectively. GPT-4o does so to a lesser but still significant extent (34%). Overall, GPT-4o demonstrates a stronger grasp of dialectal structure and vocabulary, has better visual grounding in addition to a superior cultural understanding and less generation failure.

Cross-Dialect Analysis Next, we investigate the performance differences between dialects, averaged across models. In Figure 4b, we observe that the Moroccan and Emirati dialects are considerably more challenging than Jordanian and Egyptian. The models seem to struggle with the Emirati cultural aspects the most, and the linguistic aspects of both Emirati and Moroccan, but more so on the latter. All dialects seem to exhibit a similar degree of *partial MSA use* but Jordanian and Egyptian do not pose problems in other dialectal error categories.

6 Conclusion

In this paper, we presented JEEM, a culturally-informed benchmark for VLMs that covers four diverse Arabic-speaking countries: **J**ordan, the **E**mirates, **E**gypt, and **M**orocco. By incorporating both image captioning and visual question answering, JEEM allows for a comprehensive evaluation of VLM generalization capabilities across different Arab cultural and dialectal contexts. We also included an evaluation of four Arabic VLMs: Peacock, Maya, AIN, and Palo, in addition to the multilingual GPT-4o. We included both human and automatic evaluation results to ensure reliability, and attempted to measure performance across various dimensions. Our results indicate that current models, whether general or Arabic-specific, still struggle with dialectal and cultural understanding. GPT-4o achieved the highest scores on most metrics, but there is large room for improvement in semantic dimensions like relevance and consistency. In addition, models struggle more with low-resource dialects like Emirati compared to high-resource variants like MSA and Egyptian.

Limitations

While JEEM provides a culturally diverse benchmark for VLMs, several limitations should be acknowledged. First, the dataset focuses on only four Arabic dialects, leaving out many others and thus limiting a comprehensive and inclusive evaluation. Second, although JEEM serves as a benchmark rather than a training dataset, its size remains relatively small compared to existing Western vision-language datasets. Additionally, automatic evaluation metrics such as CIDEr and BLEU may not fully capture the complexity of dialect-specific and culturally nuanced responses. While we incorporate human evaluation, it is conducted on only a subset of the data. Expanding human evaluation to a broader sample could provide a more comprehensive assessment of model performance.

Ethics Statement

Fair Job Conditions Our team of writers is based in the United Arab Emirates, Jordan, Morocco, and Egypt. Their pay rates exceed the respective hourly minimum wages. Annotation and voluntary survey results are collected and stored anonymously. Writers are informed in advance about potentially sensitive or harmful content in the images, which may be related to topics such as politics, culture, and religion.

Licensing Information The images are subject to the underlying licensing terms of Wikimedia Commons [4] and Flickr [5]. The image captions, questions, and answers are distributed under the MIT license[6].

References

Huda A. Al-muzaini, Tasniem N. Al-yahya, and Hafida Benhidour. 2018. Automatic arabic image captioning using rnn-lstm-based language model and cnn. *International Journal of Advanced Computer Science and Applications*, 9(6).

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, et al. 2024. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged Al-shaibani. 2024. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand. Association for Computational Linguistics.

Zaid Alyafeai, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating arabic nlp tasks using chatgpt models. *Preprint*, arXiv:2306.16322.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782.

Kyle Buettner and Adriana Kovashka. 2024. Quantifying the gaps between translation and native perception in training for multimodal, multilingual retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5863–5870, Miami, Florida, USA. Association for Computational Linguistics.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. Maxm: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.

Cohere. 2025. Aya vision: Expanding the worlds ai can see. Accessed: March 21, 2025.

Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59.

[4] https://wikimedia.org/Licensing/
[5] https://flickrhelp.com/creativecommons/
[6] https://opensource.org/license/mit

Obeida ElJundi, Mohamad Dhaybi, Kotaiba Mokadam, Hazem M Hajj, and Daniel C Asmar. 2020. Resources and end-to-end neural network models for arabic image captioning. In *VISIGRAPP (5: VISAPP)*, pages 233–241.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 900–913.

Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 2023. Exploring why object recognition performance degrades across income levels and geographies with factor annotations. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nizar Y. Habash. 2010. *Introduction to Arabic natural language processing*, 1 edition, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. Ain: The arabic inclusive large multimodal model. *arXiv preprint arXiv:2502.00094*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Vasu Jindal. 2017. A deep learning approach for arabic caption generation using roots-words. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi. 2018. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 402–412, Brussels, Belgium. Association for Computational Linguistics.

Sarah M Kamel, Shimaa I Hassan, and Lamiaa Elrefaei. 2023. Vaqa: Visual arabic question answering. *Arabian Journal for Science and engineering*, 48(8):10803–10823.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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 *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

LlamaTeam. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Rasha Mualla and Jafar Alkheir. 2018. Development of an arabic image description system. *Int. J. Comput. Sci. Trends Technol*, 6:205–213.

Richard E Nisbett and Takahiko Masuda. 2013. Culture and point of view. In *Biological and cultural bases of human inference*, pages 49–70. Psychology Press.

Team OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A polyglot large multimodal model for 5b people. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. 2024. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. *Preprint*, arXiv:2412.13647.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13559–13568.

Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, and Wanxiang Che. 2024. Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation. *Preprint*, arXiv:2407.01081.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. 2025. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. *Preprint*, arXiv:2503.07906.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Annotation Statistics

See Figure 5 for dialectal coverage. See Table 5 for writer profiles. See Figure 6 for tasks distribution.

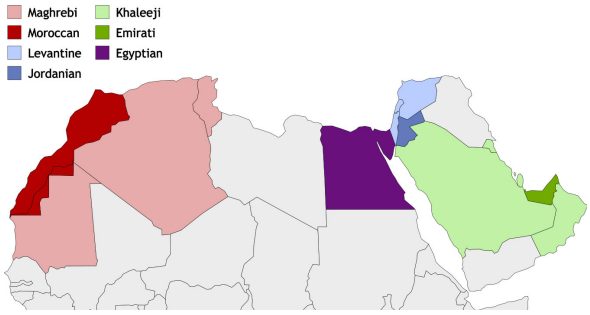


Figure 5: Dialectal coverage of JEEM. The country-level dialects used are shown in dark colors along with their respective region-level dialects in lighter color. The regional classification follows the work of Habash (2010).

B Annotation Guidelines

B.1 Image Captioning

You are presented with a photo that depicts a scene from daily life (e.g., food, clothing, homeware), social life (e.g., public transport, road signs, public ads), or urban objects from your area. Your task is to write a description of this photo in Arabic.

Steps for Writing

1. **Analyze the photo:** Identify key elements, people, objects, actions, and any relevant background details.
2. **Write the description of the photo:** The description should provide essential information. Typically, 15-25 words are sufficient. Describe everything that adds value and clarity.
3. **Explain what is behind the scenes:** If necessary, describe the context of the photo using your background knowledge (e.g., where the photo could have been taken, whether the food in the photo is special, etc.).
4. **Use everyday language:** Use ordinary informal language, but feel free to incorporate slang where appropriate.

Hints for Creating a Better Description

Your description should be detailed enough to give a clear idea of what is happening in the photo to someone who cannot see it. Try to include details that are specific to your culture or region. Here are some hints to help you:

- **Describe people, animals, objects, and key elements:** How they look and how they relate to each other in the physical space.
- **Describe interactions:** Who or what interacts with whom or what, and how they interact.
- **Include implicit details:** Add information that is not explicitly presented in the photo if it helps convey the image better. For example, if you can tell from people's attire that this is a wedding party, even though there is no visible banner stating so, mention it in the description.
- **Use precise terminology:** "Cat" is better than "animal", and "Siamese cat" is better than "cat."
- **Rely on everyday knowledge and culture, but avoid over-fantasizing:** You do not need to create a story or a plot, but you should be as precise as possible in your description.

B.2 Question Writing

You are presented with a description of a photo, but you do not have access to the photo itself. Your task is to ask five questions that will help you better understand what is happening in the photo and refine the description.

Steps for Writing

1. **Carefully read the description:** Identify parts that are unclear, ambiguous, or lacking in detail.
2. **Formulate a question:** Craft a question to clarify ambiguities or add relevant details to the photo description.
3. **Use everyday language:** Use ordinary, informal language, but feel free to incorporate slang where appropriate.

Hints for Creating Better Questions

- **Pay attention:** Do not ask for details that are already provided. For example, if the description states, "The photo shows a woman in a red dress," you should not ask, "What color is the dress in the photo?"
- **Keep questions concise:** Questions should be no longer than one sentence. There is no need to provide additional context within the question.
- **Base your questions on the description:** You can inquire about people, animals, or objects mentioned—how they look, what people are wearing, what they are doing, how they

Question	Response (%)
What gender do you identify as?	Male: 45.8, Female: 54.2, Nonbinary/Other: 0
What is your age?	20-29: 50, 30-39: 29.2, 40-49: 20.8, 50+: 0
What is your nationality?	Jordan: 37.5, Egypt: 29.2, Morocco: 20.8, UAE: 12.5
What is your native language?	Arabic: 95.8, Multiple incl. Arabic: 4.2
What is your native dialect?	Jordanian: 37.5, Egyptian: 29.2, Darija: 20.8, Emirati: 12.5
Where did you grow up? (Nearest city)	Jordan: Amman (33.3), Irbid (4.2); Morocco: Tetouan (8.4), Casablanca (8.4), Khenifra (4.2); Egypt: Cairo (8.4), Giza (4.2), Mansoura (4.2), Tanta (4.2), Damietta (4.2), Helwan (4.2); UAE: Al Ain (4.2), Abu Dhabi (4.2), Ajman (4.2)
Highest level of education?	High school: 4.2, Undergraduate: 41.7, Postgraduate: 29.2, Master's: 29.8, Doctorate: 4.2
Years of work experience?	1-3: 37.5, 4-6: 12.5, 7-9: 16.7, 10-12: 16.7, 13-15: 4.2, 16+ years: 12.5
What is your current employment status?	Not working: 8.2, Self-employed: 25, Part-time: 33.3, Full-time: 33.3

Table 5: Results of the voluntary survey of 24 respondents.

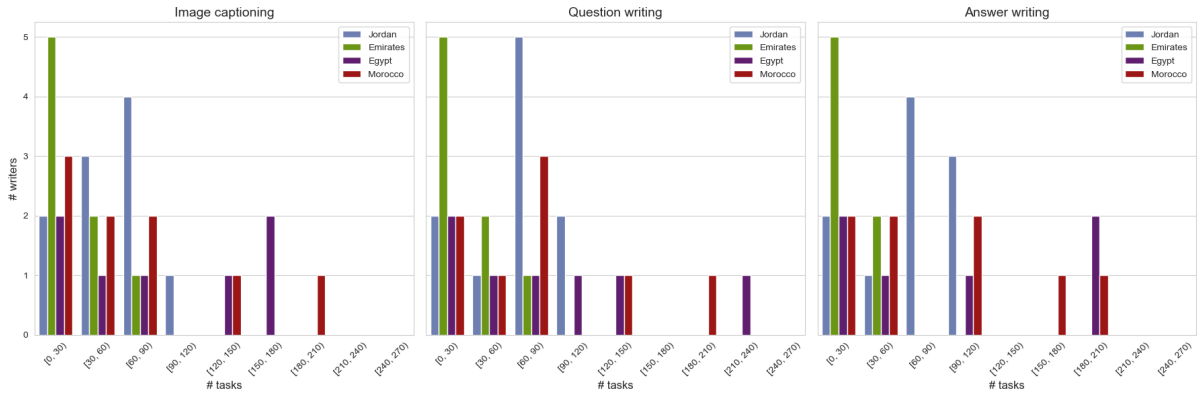


Figure 6: Distribution of annotators based on the number of tasks completed for three tasks: Image Captioning, Question Writing, and Answer Writing. Each bar represents the number of writers contributing within a given range, with colors indicating different dialects. Y-axis: number of unique writers. X-axis: the number of tasks grouped into intervals.

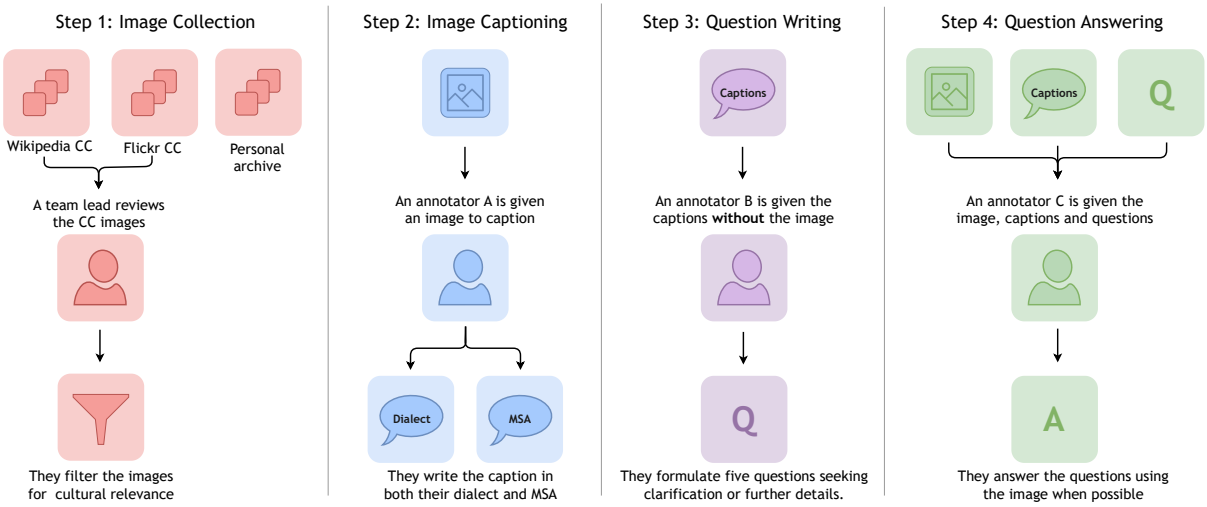


Figure 7: Data collection pipeline.

relate to each other in physical space, and how they interact.

- **Ask about background details:** Consider why people are dressed a certain way, why they are performing specific actions, or why certain objects are present.
- **Inquire about future events:** Ask what might happen next—what people will do right after the described scene, or what will happen to the objects mentioned.
- **Request emotional or aesthetic judgments:** Ask whether the photo looks nice, whether it would work as a postcard, or whether it would make a good wall print.
- **Avoid unnecessary repetition:** You do not need to repeat the exact wording from the description in your question. For example, if the description states, "The picture shows an empty street with a single car passing by," you do not have to use the word "car" in your question. Instead of asking, "What color is the car?" you can ask, "What color is it?"

B.3 Question Answering

You are presented with a photo that shows a scene from daily life (e.g., food, clothing, homeware), social life (e.g., public transport, road signs, public ads), or urban objects from your area, along with a description of this photo and five questions asking to clarify missing information from the photo. Your task is to answer the questions.

Steps for Writing

1. **Analyze the photo:** Identify key elements, people, objects, actions, and any relevant background details.
2. **Carefully read the description and the questions:** Identify what is unclear and missing in the description.
3. **Answer the questions:** Provide a clear and detailed answer based on the photo to clarify or add to its description. Aim for 2-3 sentences.
4. **Use everyday language:** Use ordinary, informal language, but feel free to use slang words where necessary.
5. **Revise, Edit, Submit.**

Hints for Creating Better Answers

- **Take your time to carefully look over the photo:** Pay attention even to the smallest details before answering each question.
- **Base your answer on the photo or your cultural knowledge:** You do not need to create a story or explanation if it cannot be gathered from the photo.
- **If the question cannot be answered:** If something is not clear from the photo or your cultural knowledge, choose the option `<Cannot tell from the picture>`.
- **If details are already mentioned in the description:** You may simply copy the answer from there if the question asks for details that have already been stated.

B.4 Human Evaluation of Image Captioning

You are presented with an image and an image caption — a short text that describes the content of the image. You need to look closely at the image, read its caption and evaluate the caption according to the following five criteria. Evaluate each criterion on a scale from 1 to 5, where 1 means very bad, 3 means neutral, and 5 means excellent. Be lenient; when in doubt, don't be afraid to give a high score.
Consistency: Does the caption match what is actually shown in the image? It should avoid adding details that are not visible.
Relevance: Does the caption mention the most important elements in the image? It should focus on the main subjects without omitting key details.
Fluency: Evaluate how naturally and smoothly the text reads. Consider clarity, word choice, and overall ease of understanding. A fluent text should be easy to read, free of language errors, and sound natural.
Dialect authenticity: How well does the caption represent the spoken dialect in your country? Does it use words and phrases that people in your country commonly would use?
Note that fluency and dialectal language are not the same. Fluency evaluates how natural and correct the caption is, while dialectal language assesses how dialectal or regional the caption sounds. A caption might be non-fluent but still dialectal or attempting to sound dialectal.

The `JEEM` dataset consists of 2,196 instances. Each instance includes: *i*) 2 image captions *ii*) 5 questions *iii*) 5 corresponding answers. This results in: *i*) 4,392 captions *ii*) 10,980 questions *iii*) 10,980 answers.

Time Estimation for Dataset Creation Estimated time to create one instance: *i*) Image captioning: 10 minutes *ii*) Question writing: 7 minutes

iii) Answer writing: 7 minutes *iv*) Validation: 10 minutes.

Total time per instance: 34 minutes

Total time for 2,196 instances: $2,196 \times 34$ minutes $= 74,664$ minutes $\approx 1,245$ hours

Considering a rework rate of approximately 30%, the actual time spent is: $1,245$ hours $\times 1.3 \approx 1,618$ hours

Human Evaluation of Image Captioning Human evaluation was conducted across three dialects. The setup involved: *i*) 100 images *ii*) caption generated by 6 models and 1 human writer *iii*) 3-way overlap.

Total annotations per dialect: $100 \times (6+1) \times 3 = 2,100$

Across three dialects: $3 \times 2,100 = 6,300$ annotations

Due to slower annotation speed in the Emirati group, the evaluation procedure is modified: *i*) 50 images $\times$ (6 models + 1 human) $\times$ 1 overlap = 350 annotations *ii*) extra 50 random images $\times$ 2 overlap = +50 annotations.

Total annotations in the UAE subset: $350 + 50 = 400$ annotations

Time Estimation for Human Evaluation Each caption evaluation takes approximately 5 minutes.

Standard Evaluation Time: $6,300 \times 5 = 31,500$ minutes $= 525$ hours

UAE Subset Evaluation Time: $350 \times 5 = 1,750$ minutes $= 29.2$ hours

Summary *i*) Total human evaluations for image captioning: 6,650 annotations *ii*) Total evaluation time: 525 + 29.2 = **554.2 hours**

C Topic Categories

The topic categorization is presented in Tables 6 and 7, which systematically organize the dataset's content into distinct thematic groups. While all categories demonstrate significant representation, the Arts & Culture classification warrants particular attention due to its exceptional diversity despite comprising a relatively smaller proportion of the overall dataset. This category encompasses a rich variety of subdomains including performing arts (traditional dance and music), visual arts (henna designs and pottery decoration), culinary traditions (traditional food preparation), and heritage crafts (ceramics and textile arts).

D Cultural Aspects

The dataset reveals significant overlap in cultural concepts across different dialects. However, we also observed frequent instances where annotators misidentified visually similar items, as illustrated in Figure 8. Another notable example involves white spherical objects that Emirati annotators identified as cheese, while Jordanian annotators labeled them as Jameed (a traditional dried yogurt used for cooking Mansaf). These cases highlight the challenges in cross-cultural visual identification, particularly with regionally specific items.

E Prompts

The prompts used for evaluation can be seen in Figures 9, 10, 11. In all cases, the evaluation is based on four criteria: Consistency, Relevance, Fluency, and Dialect Authenticity, following a structured format and a five-point rating scale. For MSA, only the first three criteria (Consistency, Relevance, and Fluency) were included in the prompt, while Dialect Authenticity was omitted.

F Human Evaluation of Ground-truth Reference

Table 8 reports human evaluation scores for the reference captions across all dialects. These serve as an upper bound for model performance and provide a useful point of comparison for evaluating the quality of generated captions.

	Con	Rel	Flu	DAuth
MSA	4.59	4.57	4.68	4.88
JO	4.59	4.57	4.68	4.88
AE	4.48	4.72	4.92	4.94
EG	4.81	4.74	4.78	4.93
MA	4.91	4.69	4.90	4.96

Table 8: Human evaluation scores across all dialect subsets of JEEM. Human captions consistently achieve the highest scores across all criteria.

G Reference-Only and HalFScore Evaluation

In addition to our main evaluation setup, we conduct two supplementary experiments to further probe model behavior—specifically focusing on hallucination and the role of visual grounding.

Places	سوق، مسجد، مدينة، الصحراء، قرية، حديقة، كورنيش، قرية جبلية، مدينة ساحلية، منطقة سكنية، منطقة جبلية، مطار، ميناء، مركز تجاري، مركز الخدمات الطبية، مكتبة الجامعة، مقبرة، مبنى حكومي، مبنى البريد، محطة وقود
Translation	Market, Mosque, City, Desert, Village, Park, Corniche, Mountain Village, Coastal City, Residential Area, Mountain Area, Airport, Port, Mall, Medical Center, University Library, Cemetery, Government Building, Post Office, Gas Station
Celebrations	عيد الميلاد، احتفال، حفل زفاف، حفل توزيع الجوائز، مهرجان، مهرجان فولكلوري، اجتماع، ندوة، محاضرة، عرض موسيقي، عرض تقديمي، تظاهرة، حملة توعية، حملة تطوعية
Translation	Christmas, Celebration, Wedding, Award Ceremony, Festival, Folk Festival, Meeting, Seminar, Lecture, Music Performance, Presentation, Demonstration, Awareness Campaign, Volunteer Campaign
Arts & Culture	رقص تقليدي، عرض موسيقي، فيلم، حفلة، موسيقية، ورشة عمل فنية، تحضير الطعام التقليدي، تحضير الحلويات، الطبخ التقليدي، الفن التقليدي، التراث، الأسواق التقليدية، الحرف التقليدية، منتجات خزفية، منتوجات تقليدية، تذكارات نقش الحناء، نقش البلاط، تلوين المزهريات
Translation	Traditional Dance, Music Performance, Film, Party, Musical, Art Workshop, Traditional Food Preparation, Sweets Preparation, Traditional Cooking, Traditional Art, Heritage, Traditional Markets, Traditional Crafts, Ceramic Products, Traditional Products, Souvenirs, Henna Art, Tile Art, Pottery Decoration
Nature	غابة، طبيعة، وادي رم، وادي دادس، شجرة نخيل، غروب الشمس، منظر طبيعي، واحة صحراوية، جبال، الصحراء، الطبيعة المغربية، الشتاء، الربيع، الزقاق المغربي، الحديقة، حديقة عامة، حديقة حيوانات، حديقة ماجوريل، الطبيعة المغربية
Translation	Forest, Nature, Wadi Rum, Dades Valley, Palm Tree, Sunset, Scenic View, Desert Oasis, Mountains, Desert, Moroccan Nature, Winter, Spring, Moroccan Alley, Garden, Public Park, Zoo, Majorelle Garden, Moroccan Nature
Education	مدرسة، الثانوية التأهيلية الحسن الثاني، تعليم اللغة الفرنسية، تدريب المعلمين، محل خياطة، مخيم كشافة، ورشة عمل، مكتبة الجامعة، المدرسة، فصل دراسي
Translation	School, Hassan II High School, French Language Education, Teacher Training, Tailor Shop, Scout Camp, Workshop, University Library, School, Classroom
Transport	محطة حافلات، محطة ترامواي، محطة سيارات الأجرة، سيارة أجرة، طريق، طريق سيار، طريق نائية، سيارات الأجرة، ترامواي، قطار، القطارات، موقف سيارات، موقف سيارات الأجرة، شاحنة، سيارة قديمة، توك توك، حافلة
Translation	Bus Station, Tram Station, Taxi Stand, Taxi, Road, Highway, Remote Road, Taxis, Tram, Train, Trains, Parking Lot, Taxi Parking, Truck, Old Car, Tuk Tuk, Bus
Characters	الملك محمد السادس، الأميرة للا سلمى، شخصيات سياسية، مقدم، رجل مسن، نساء من شمال المغرب، رجل، فتيات، أطفال، عامل بناء، بائع، بائع متجول
Translation	King Mohammed VI, Princess Lalla Salma, Political Figures, Presenter, Elderly Man, Women from Northern Morocco, Man, Girls, Children, Construction Worker, Seller, Street Vendor

Table 6: Topic Categories, Part 1.

F&B	طعام، فلافل، عصير البرتقال، شاي، شاي مغربي، خبز، الكبسة، الكسكس، مخللات، تحلية، عصائر طبيعية، وجبة طعام، فواكه، زيت الأركان، أملو، هريس، فواكه جافة، خبز مقلي، الكنافة، الخبز، الطعام التقليدي، المأكولات المغربية
Translation	Food, Falafel, Orange Juice, Tea, Moroccan Tea, Bread, Kabsa, Couscous, Pickles, Dessert, Fresh Juices, Meal, Fruits, Argan Oil, Amlou, Porridge, Dried Fruits, Fried Bread, Kunafa, Bread, Traditional Food, Moroccan Cuisine
Sports	رياضة، كرة القدم، رياضات قتالية، فروسية، سباق السيارات، ركوب الجمل، المنتخب المغربي، المنتخب المغربي لكرة القدم، مباراة رياضية
Translation	Sports, Football, Martial Arts, Equestrianism, Car Racing, Camel Riding, Moroccan National Team, Moroccan National Football Team, Sports Match
Trade	التجارة، محل زيتون، محل مكسرات، محل بيع المنتجات التقليدية، محل بيع التمر، محل خياطة، محل بيع الفاكهة، محل الأعشاب والتوابل، محل بيع الفواكه، محل بيع الخبز، محل بيع الخضار، محل بيع المكسرات، محل بيع الحلوى
Translation	Commerce, Olive Shop, Nut Shop, Traditional Products Shop, Date Shop, Tailor Shop, Fruit Shop, Herbs and Spices Shop, Fruit Shop, Bread Shop, Vegetable Shop, Nut Shop, Sweet Shop
Technology	معدات طبية، تسوق، عرض مشروع، عرض تقديمي، عرض جوي، تكنولوجيا، اتصالات المغرب، تكنولوجيا المعلومات
Translation	Medical Equipment, Shopping, Project Presentation, Presentation, Aerial Show, Technology, Maroc Telecom, Information Technology
Games	ألعاب، لعبة الدومينو، ألعاب أطفال، ألعاب تقليدية، ألعاب رياضية
Translation	Games, Dominoes, Children's Games, Traditional Games, Sports Games
Others	الخطوط الملكية المغربية، الشرطة، الحرس الملكي، الجيش، الاستعمار الفرنسي في المغرب، دبلوماسية، البرلمان المغربي، المديرية العامة للأمن الوطني، الهيئة الوطنية للسلامة المرورية، الهيئة الوطنية للوقاية من حوادث السير
Translation	Royal Air Maroc, Police, Royal Guard, Army, French Colonization in Morocco, Diplomacy, Moroccan Parliament, Directorate General of National Security, National Road Safety Authority, National Traffic Accident Prevention Authority

Table 7: Topic Categories, Part 2.

Jordanian

هاي صورة فيها طبق حلو تقليدي محطوط بصحن قزاز كبير وشفاف، ومغلف من تحت بالبلاستيك عشان يضل نضيف، والحلو مزين بانواع مختلفة من المكسرات زي **اللوز** المبشور **والفستق**، وعلى الوجه في حبة جوز، والحلو لونه غامقة على الاغلب معمول **بالكراوية** او **الدبس**.

This image contains a traditional dessert in a big and transparent glass plate, and it's covered on the bottom with plastic so that it remains clean, and the dessert is decorated with different nuts such as grated **almonds** and **pistachios** and on the surface is a walnut, and the dessert's color is dark and it is likely made of **karawya** or **dibs** (fruit syrup).

Emirati

طاسة زجاجية فيها **حلوى عمانية**، وعليها مكسرات منوعة. عدالها غطا الطاسة. شكلها الطاسة بديدة وتوهم فاتحينها، لأن أغلب الحلوى محد هابشنها، وبعده جزء من الطاسة مغطاي بنايلون. الطاسة محطوطه على باركيه بني.

A glass bowl with **Omani halwa** topped with mixed nuts. Next to it is the bowl's lid. The bowl looks new and just opened because most of the halwa has not been touched, and part of the bowl is still covered with plastic wrap. The bowl is placed on a brown wooden floor.

Egyptian

صورة **لبودنج** من لونه اكيد بودنج **شيكولاتة** عليها قطعة عين جمل وشرايح **صنوبر** صغيرة على الوش في طبق حلويات ازاز عميق وصغير وله غطا ازاز مفتوح ومحطوط ساند على الطبق وعليه غلاف بلاستيك شفاف لسة مفتوح نازل على نص الطبق. الطبق محطوط على الارض اللي معمول من الباركيه شكل الخشب.

A picture of a **pudding**, that is from its color definitely **chocolate** pudding, it has a walnut and slices of **pine seeds** on the surface in a deep, glass desserts plate it has a glass lid that is open and leaning on the plate and has an opened transparent plastic cover reaching only half the plate. The plate is on the floor which is made of parquet looking like wood.

Moroccan

هاد التصويرة كتبان فيها واحد الطاسة صغيرة ديال الجاج مغلفة بسولوفان من التحت وفهاد الطاسة كيبان فيها تحلية كحلة يمكن **شكلاط**، فالزاج ديال الطاسة كنشوفو شي حاجة بحال **كراميل** وهاد التحلية مزوقة **بالكوكو وبيسطاش** وواحد الكركاعة فالوسط.

This image shows a small glass plate covered in cellophane from the bottom, and in this plate we see a black dessert that could be **chocolate,** on the plate's side we see something that looks like **caramel** and this dessert is decorated with **coconut** and **pistachios** and a walnut in the middle.

Figure 8: Image of a Omani Halwa (image sourced from the Emirati set) shared with annotators across all dialects. The Jordanian, Egyptian and Moroccan captions demonstrate an incorrect identification of the dessert and its components.

You are an expert evaluator assessing the quality of an Arabic image caption. You will be given an image, a reference caption, and a caption to evaluate. Your task is to carefully analyze all three and evaluate the given caption based on four criteria: **Consistency, Relevance, Fluency, and Dialect Authenticity.**
Evaluate each criterion on a scale from 1 to 5, where 1 means very bad, 3 means neutral, and 5 means excellent.
Consistency: Does the caption match what is actually shown in the image? It should avoid adding details that are not visible.
Relevance: Does the caption mention the most important elements in the image? It should focus on the main subjects without omitting key details.
Fluency: Evaluate how naturally and smoothly the text reads. Consider clarity, word choice, and overall ease of understanding. A fluent text should be easy to read, free of language errors, and sound natural.
Dialect Authenticity: How well does the caption represent the spoken dialect in `{country}`? Does it use words and phrases that people in this country commonly would use?
Reference Caption: `{reference}`
Generated Caption: `{generated}`
Output Format (do not add any additional information):
Consistency: X/5
Relevance: X/5
Fluency: X/5
Dialect Authenticity: X/5

Figure 9: Evaluation prompt for assessing Arabic image captions **using both the image and the reference caption**.

You are an expert evaluator assessing the quality of an Arabic image caption. You will be given a reference caption and a caption to evaluate. Your task is to carefully compare the evaluated caption to the reference caption and assess it based on four criteria: **Consistency, Relevance, Fluency, and Dialect Authenticity.**
Evaluate each criterion on a scale from 1 to 5, where 1 means very bad, 3 means neutral, and 5 means excellent.
Consistency: Does the evaluated caption match the reference caption in meaning and key details? It should avoid adding information that is not present in the reference caption or contradicting its content.
Relevance: Does the evaluated caption mention the most important elements described in the reference caption? It should focus on the main subjects without omitting key details.
Fluency: Evaluate how naturally and smoothly the text reads. Consider clarity, word choice, and overall ease of understanding. A fluent text should be easy to read, free of language errors, and sound natural.
Dialect Authenticity: Check how well the caption represents the dialect spoken in `{country}`. Does it use words and phrases that people in this country commonly use?
Reference Caption: `{reference}`
Generated Caption: `{generated}`
Output Format (do not add any additional information):
[same as above]

Figure 10: Evaluation prompt for assessing Arabic image captions **using only the reference caption**.

You are an expert evaluator assessing the quality of an answer to a question in dialectal Arabic. You will be given an image, a question, a reference answer, and an answer to evaluate. Your task is to carefully analyze all four and evaluate the given answer based on four criteria: **Consistency, Relevance, Fluency, and Dialect Authenticity.**
Evaluate each criterion on a scale from 1 to 5, where 1 means very bad, 3 means neutral, and 5 means excellent.
Consistency: Does the answer correctly address the question while accurately describing the content of the image? It should avoid adding details that are not visible.
Relevance: Does the answer provide the most important details necessary to respond to the question based on the image? It should focus on the main subjects without omitting key details.
Fluency: Evaluate how naturally and smoothly the answer reads. Consider clarity, word choice, and overall ease of understanding. A fluent answer should be easy to read, free of language errors, and sound natural.
Dialect Authenticity: How well does the answer represent the spoken dialect in `{country}`? Does it use words and phrases that people in this country commonly would use?
Question: `{question}`
Reference Answer: `{reference}`
Generated Answer: `{generated}`
Output Format (do not add any additional information and do not add explanations at the end of the evaluation):
[same as above]

Figure 11: Evaluation prompt for assessing answers to questions in dialectal Arabic using an image, a question, and a reference answer.

GPT Evaluation (Reference Only). To isolate the contribution of visual input in GPT-based scoring, we repeat our four-criteria evaluation (Consistency, Relevance, Fluency, and Dialect Authenticity), but this time providing GPT-4o only with the reference caption and the model-generated caption—excluding the image. This setting allows us to examine how reliably GPT-4o can assess captions based solely on textual alignment. As expected, it performs slightly worse at identifying visually grounded content, but remains effective in evaluating fluency and general semantic coherence. Dialect judgments may also degrade in this setup due to lack of visual context.

HalFScore. We also evaluate captions using HalFScore (Chen et al., 2025), a recent metric developed to assess hallucination and omission in image captioning. HalFScore parses each caption into object–attribute–relation triplets using GPT-4o and compares them to triplets extracted from the reference. It then computes precision, recall, and F_1 at the scene-graph level, providing a complementary signal to metrics like DCScore. Although HalFScore was originally proposed for longer, multi-sentence captions, we find it still useful for single-sentence outputs in highlighting factual mismatches.

However, we observe a key limitation in how HalFScore handles omissions: if a model misses key objects that form the basis of several relations, those dependent relations are not double-counted as omissions. As a result, captions that omit all semantically critical content may still receive moderately high recall scores. This partly explains the metric's weaker correlation with human relevance ratings, particularly for captions judged extremely uninformative by annotators.

Results. Table 9 presents the results from both evaluation settings. These metrics are intended as complementary analyses and are not used to draw primary conclusions in the main paper. Full implementation details and limitations are discussed in Appendix G.

H Maya and Peacock Model Performance

Table 10 provides full evaluation results for Maya and Peacock, the two lowest-performing models in our benchmark. These were excluded from the main results table for clarity but are included here for completeness.

I Implementation Details for DCSCORE

DCSCORE (Ye et al., 2025) is designed to evaluate detailed image captioning by comparing structured units of meaning, called Primitive Information Units (PIUs), between generated and reference captions. The official implementation is provided in the DeCapBench repository.[7]

Following a three-stage pipeline, PIU decomposition, image-grounded verification, and semantic matching, DCSCORE computes multiple evaluation variants:

- **Precision**, **Recall**, and **F1-score**, computed over all verified PIUs;

[7] https://github.com/MAGAer13/DeCapBench

	Model	GPT Eval (Reference Only)				HalFScore		
		Con	Rel	Flu	DAuth	Prec.	Rec.	F1
JO	AIN	1.6	1.7	4.2	2.7	45.0	46.1	44.2
	AyaV	1.9	2.1	4.1	3.1	48.4	**47.9**	46.3
	Maya	1.5	1.6	4.0	2.7	**49.9**	43.9	45.0
	Palo	1.7	1.8	4.1	2.7	49.8	44.7	45.0
	Peacock	1.4	1.5	3.4	2.5	42.7	42.7	40.2
	GPT-4o	**2.3**	**2.7**	**4.6**	**3.7**	49.4	47.4	**46.8**
AE	AIN	1.5	1.6	4.2	2.1	44.6	43.4	42.7
	AyaV	1.7	1.8	4.2	2.3	45.2	41.3	41.3
	Maya	1.5	1.6	4.2	2.1	**51.5**	42.0	**44.7**
	Palo	1.6	1.7	4.1	1.9	50.2	41.2	43.1
	Peacock	1.2	1.3	3.8	1.7	45.5	41.8	**41.3**
	GPT-4o	**2.0**	**2.3**	**4.5**	**2.6**	46.6	**44.1**	43.7
EG	AIN	1.4	1.5	4.1	2.4	46.7	42.2	42.9
	AyaV	1.7	1.9	4.0	4.0	47.7	44.2	44.9
	Maya	1.3	1.4	4.0	2.3	48.8	40.6	43.4
	Palo	1.5	1.6	4.0	2.3	**49.6**	41.7	44.5
	Peacock	1.2	1.2	3.8	2.3	43.7	39.7	40.3
	GPT-4o	**2.0**	**2.3**	**4.3**	**4.3**	49.1	**44.4**	**45.6**
MA	AIN	1.4	1.5	4.1	1.3	44.7	40.7	40.2
	AyaV	1.8	2.0	3.9	4.2	46.5	39.3	40.7
	Maya	1.5	1.6	4.0	1.2	49.1	39.0	42.0
	Palo	1.5	1.6	4.0	1.2	**51.6**	**40.9**	**44.0**
	Peacock	1.3	1.4	3.4	1.9	45.9	36.5	38.8
	GPT-4o	**2.1**	**2.5**	**4.3**	**4.5**	47.1	40.3	41.1
	τ_c	25.9	29.2	16.9	44.3	7.1	4.8	7.2

Table 9: **Supplementary Evaluation.** GPT-4o evaluation in the *reference-only* setting and HalFScore for all models across four Arabic dialects. All results are computed on the 350-image evaluation subset. HalFScore reports hallucination-adjusted precision, recall, and F_1 over scene elements. τ_c denotes Kendall's Tau-c correlation with human rankings.

- **Precision$_{\text{relevant}}$**, **Recall$_{\text{relevant}}$**, and **F1$_{\text{relevant}}$**, computed over a subset of "relevant" PIUs (those conveying core visual content);
- **DCScore**, defined as the average of the F1-score and $F1_{\text{relevant}}$:

$$\text{DCSCORE} = \frac{F1 + F1_{\text{relevant}}}{2}.$$

In our evaluation, we report only the standard **F1-score** in the main results table, as it showed the highest correlation with human judgments across the 350-image evaluation set. Both $F1_{\text{relevant}}$ and the averaged DCScore yielded weaker alignment with human ratings. All results are based on the official implementation.

J Error Analysis

See Figure 14 for model error patterns separated by dialect. See Table 11 for error types.

	Model	Traditional Metrics				GPT-4-as-a-Judge*				DCScore*			Human Eval*			
		B	C	R	BSc	Con	Rel	Flu	DAuth	Precision	Recall	F1-score	Con	Rel	Flu	DAuth
MSA	Maya	4.25	1.79	9.47	90.35	2.30	2.39	4.05	-	55.26	38.93	44.77	-	-	-	-
	Peacock	2.08	1.51	7.18	84.24	1.92	1.96	3.34	-	58.68	22.60	31.46	-	-	-	-
JO	Maya	1.91	0.49	6.44	90.16	2.45	2.52	4.12	2.60	55.63	42.33	47.24	2.49	2.58	3.71	1.35
	Peacock	1.55	1.88	5.91	83.57	2.32	2.36	3.56	2.52	61.06	23.98	33.14	2.06	2.09	2.83	1.14
AE	Maya	1.55	0.36	5.98	89.43	2.44	2.42	4.20	2.04	53.56	43.22	46.70	2.06	1.84	2.20	2.00
	Peacock	1.23	0.95	4.09	79.09	2.04	2.06	3.86	1.82	58.74	23.44	32.08	1.73	2.28	2.84	1.00
EG	Maya	2.16	0.49	6.67	90.82	2.04	2.08	4.03	2.40	51.19	34.85	40.61	2.93	3.28	4.32	1.34
	Peacock	0.86	0.54	4.50	81.88	1.88	1.82	3.87	2.43	63.20	20.43	29.65	3.05	2.43	3.47	1.37
MA	Maya	1.06	0.37	3.79	88.85	2.37	2.55	4.19	1.66	56.65	45.58	49.69	3.17	3.41	4.37	1.44
	Peacock	0.51	0.40	2.55	79.88	2.18	2.11	3.44	1.99	68.56	27.00	37.16	3.07	2.51	3.95	1.00

Table 10: Performance of Maya and Peacock across all dialect subsets of JEEM. These are the two lowest-performing models in the benchmark.

Category	Type	Description
Dialectal	MSA use, partial	*Use of MSA words when dialectal alternatives exist.* مشهد رائع ديال غروب الشمس فوق المياه.
	MSA use, complete	*Use of MSA in the entire text.* الصورة تظهر مشهدًا في الشارع حيث يتم ركن سيارتين أمام مبنى.
	Made-up words	*Non-existent words that resemble dialect.* الأحمر، الأصفر، والأزرق كالت وحدات ملونة بألوان زاهية.
	Dialect mixing	*Multiple dialects mixed in one text.* الصورة دي كتعرض شجرة النخيل
	Syntactic dysfluency	*Awkward or incorrect sentence structure.* الجو حماسي بزاف، والسماء صافية، والجو كيفتح
Cultural	Underspecified CI	*Vague or imprecise mention of a cultural item.* آلة موسيقية instead of آلة الجمبري
	Misinterpretation of CI	*Misattributing a cultural item to a different culture.* تصوير يصور شخصًا يرتدي قبعة ملونة وملابس تقليدية مكسيكية
	Prompting bias	*Matches content to prompt's region, even if incorrect.*
Visual	Hallucination	*Mention of items not present in the image.* On an image with no birds: بالإضافة إلى ذلك، هناك عدة طيور
	Wrong count	*Incorrect number of items mentioned.* On an image of twelve men: رجلان يركبان على ظهور خيل سوداء
Generation Failure	Irrelevant info.	*Off-topic or unrelated content.* هاد الشجرة معروفة بفوائدها الصحية، وكتستخدم بكثرة للطعام والزيت
	Incomplete	*Ends generation abruptly.* في أحد شوارع المدينة المزدحمة
	Degeneration	*Repetitive wording or looping phrases.* هاد الصورة فيها جمل وكأنه كيف كيف كيف. وكأنه كيف كيف كيف....

Table 11: Taxonomy of the error categories and types. Examples are illustrated in *italics*. CI: cultural item.

You are an expert evaluator assessing the type of a question in dialectal Arabic. You will be given a question and must determine its type based on the following classification:
Classification:

وصف: أسئلة تهدف إلى الحصول على تفاصيل أو شرح عن موضوع أو حالة معينة.

عد: أسئلة تتعلق بعدد الأشياء أو الكميات أو تكرار حدوث شيء معين.

تحقق: أسئلة تهدف إلى التحقق من صحة أو خطأ معلومة أو حقيقة معينة.

تصنيفي: أسئلة تهدف إلى تصنيف أو تقسيم شيء ما إلى مجموعات أو أنواع أو فئات محددة.

Examples:

سؤال: شو لون عباية الحرمة اللي فالصورة؟
النوع: وصف

سؤال: الناس اللي على السقف لابسين إيه؟
النوع: وصف

سؤال: كم عدد البلدان التي تحدث فيها هذه الظاهرة؟
النوع: عد

سؤال: مبين شو اسم المطعم فالصورة؟
النوع: تحقق

سؤال: هل اليهال من نفس الأعمار؟
النوع: تحقق

سؤال: الشجر الي فالصورة شجر زينة ولا شجر مثمر؟
النوع: تصنيفي

سؤال: هاد الناس اللي فكوزينة واش رجال ولا عيالات؟
النوع: تصنيفي

Task:

السؤال:

```
{question}
```

النوع:

```
{type}
```

Figure 12: Question Type Identification Prompt. The task is to determine the type of a given question in dialectal Arabic based on a predefined classification.

You are an expert evaluator identifying the main topic of a description in Arabic. You will be given a description and must determine its main topic based on the following process:
Process:

الخطوة الأولى: قم بتعيين موضوع رئيسي لكل وصف باللغة العربية الفصحى.

الخطوة الثانية: قم بتجميع المواضيع المُحددة في فئات نهائية بناءً على التشابه بينها.

Examples:

الوصف: صورة لنخلة تظهر من الأعلى، حيث يغطّيها الكثير من الأوراق الخضراء، وتوجد بها ثمار بلح أخضر صغير لم ينضج بعد. هناك بعض الخوص يخرج من جذع النخلة وأوراقها. تم التقاط الصورة من زاوية منخفضة وقريبة من قمة النخلة.
الموضوع: طبيعة

الوصف: صورة لساحة مكان تشبه المول التجاري. الساحة مغطاة بأرضية سيراميك كبيرة بلون متدرج من البيج إلى البني. المكان عبارة عن مبنى مكون من طابق واحد، لونه بين البيج والرمادي، وجميع نوافذه من الزجاج. توجد في المبنى فواصل طويلة باللون الأحمر والأزرق والأصفر. في الساحة جزء مغطى بالزرع، وأغلب المساحة مظللة بشيء يشبه التندة. كما يوجد مجسم يشبه الساقية، بالإضافة إلى وجود أشخاص وأطفال في المكان.
الموضوع: أماكن

Task:

الوصف:

```
{caption}
```

الموضوع:

```
{topic}
```

Figure 13: Topic Identification Prompt. The task is to determine the main topic of a given caption in Arabic.

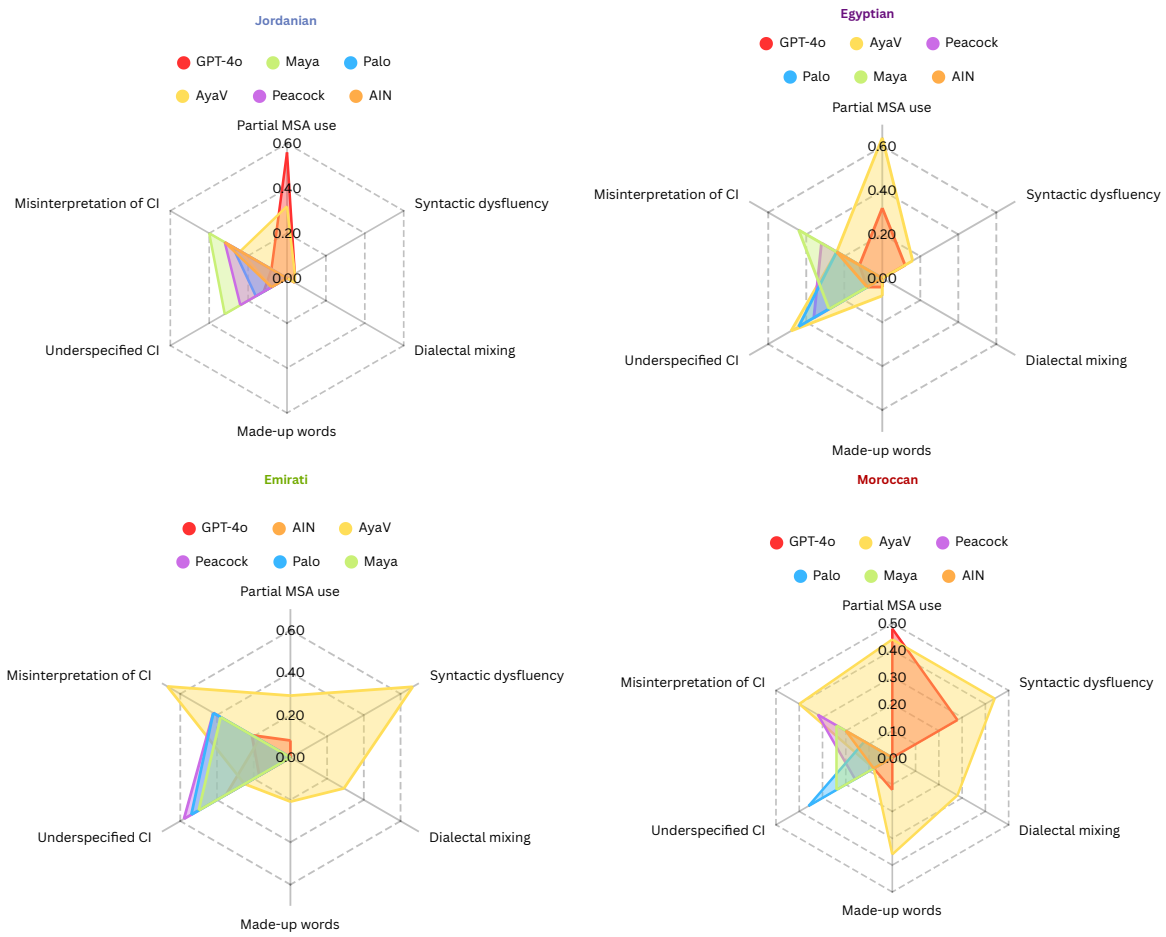


Figure 14: Model error patterns separated by dialect. CI: Cultural Item.